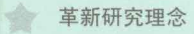


格致方法 · 定量研究系列　吴晓刚　主编

# 分析复杂调查数据(第二版)

[美] 李殷尚 (Eun Sul Lee)
罗纳德 · N. 福索佛 (Ronald N. Forthofer) 著
张卓妮 译



革新研究理念

丰富研究工具

最权威、最前沿的定量研究方法指南

23

格致出版社 上海人民出版社




## 格致方法·定量研究系列

1. 社会统计的数学基础
2. 理解回归假设
3. 虚拟变量回归
4. 多元回归中的交互作用
5. 回归诊断简介
6. 现代稳健回归方法
7. 固定效应回归模型
8. 用面板数据做因果分析
9. 多层次模型
10. 分位数回归模型
11. 空间回归模型
12. 删截、选择性样本及截断数据的回归模型
13. 应用logistic回归分析（第二版）
14. logit与probit：次序模型和多类别模型
15. 定序因变量的logistic回归模型
16. 对数线性模型
17. 流动表分析
18. 关联模型
19. 中介作用分析
20. 因子分析：统计方法与应用问题
21. 非递归因果模型
22. 评估不平等
23. 分析复杂调查数据（第二版）
24. 分析重复调查数据
25. 世代分析
26. 纵贯研究
27. 多元时间序列模型
28. 潜变量增长曲线模型
29. 缺失数据
30. 社会网络分析
31. 广义线性模型导论
32. 基于行动者的模型
33. 基于布尔代数的比较法导论
34. 微分方程：一种建模方法
35. 模糊集合理论在社会科学中的应用
36. 图形代数
37. 项目功能差异

上架建议：社会研究方法
ISBN 978-7-5432-2121-5


定价: 15.00元
易文网: www.ewen.cc
格致网: www.hibooks.cn

格致方法·定量研究系列 吴晓刚 主编

# 分析复杂调查数据（第二版）

[美] 李殷尚(Eun Sul Lee)
罗纳德·N. 福索佛(Ronald N.Forthofer) 著
张卓妮 译

SAGE Publications ,Inc.

格致出版社 上海人民出版社

**图书在版编目(CIP)数据**

分析复杂调查数据：第2版/(美)李殷禼(Lee，E.S.)，(美)福索佛(Forthofer，R.N.)著；张卓妮译．—上海：格致出版社：上海人民出版社，2012
(格致方法·定量研究系列)
ISBN 978-7-5432-2121-5

Ⅰ．①分… Ⅱ．①李… ②福… ③张… Ⅲ．①抽样调查统计-研究 Ⅳ．①C811

中国版本图书馆 CIP 数据核字(2012)第129136号

责任编辑　王亚丽

---

格致方法·定量研究系列
**分析复杂调查数据(第二版)**
[美]李殷禼　罗纳德·N.福索佛 著
张卓妮 译

---

**出　版**　世纪出版集团 www.ewen.cc　格致出版社 www.hibooks.cn　上海人民出版社
(200001　上海福建中路193号24层)

格致出版
编辑部热线 021-63914988
市场部热线 021-63914081

**发　行**　世纪出版集团发行中心
**印　刷**　浙江临安曙光印务有限公司
**开　本**　920×1168毫米 1/32
**印　张**　5.25
**字　数**　103,000
**版　次**　2012年7月第1版
**印　次**　2012年7月第1次印刷
ISBN 978-7-5432-2121-5/C·73
**定　价**　15.00元

# 出版说明

由香港科技大学社会科学部吴晓刚教授主编的"格致方法·定量研究系列"丛书，精选了世界著名的SAGE出版社定量社会科学研究丛书中的35种，翻译成中文，集结成八册，于2011年出版。这八册书分别是：《线性回归分析基础》、《高级回归分析》、《广义线性模型》、《纵贯数据分析》、《因果关系模型》、《社会科学中的数理基础及应用》、《数据分析方法五种》和《列表数据分析》。这套丛书自出版以来，受到广大读者特别是年轻一代社会科学工作者的欢迎，他们针对丛书的内容和翻译都提出了很多中肯的建议。我们对此表示衷心的感谢。

基于读者的热烈反馈，同时也为了向广大读者提供更多的方便和选择，我们将该丛书以单行本的形式再次出版发行。在此过程中，主编和译者对已出版的书做了必要的修订和校正，还新增加了两个品种。此外，曾东林、许多多、范新光、李忠路协助主编参加了校订。今后我们将继续与SAGE出版社合作，陆续推出新的品种。我们希望本丛书单行本的出版能为推动国内社会科学定量研究的教学和研究作出一点贡献。

# 总序

往事如烟，光阴如梭。转眼间，出国已然十年有余。1996年赴美留学，最初选择的主攻方向是比较历史社会学，研究的兴趣是中国的制度变迁问题。以我以前在国内所受的学术训练，基本是看不上定量研究的。一方面，我们倾向于研究大问题，不喜欢纠缠于细枝末节。国内一位老师的话给我的印象很深，大致是说：如果你看到一堵墙就要倒了，还用得着纠缠于那堵墙的倾斜角度究竟是几度吗？所以，很多研究都是大而化之，只要说得通即可。另一方面，国内（十年前）的统计教学，总的来说与社会研究中的实际问题是相脱节的。结果是，很多原先对定量研究感兴趣的学生在学完统计之后，依旧无从下手，逐渐失去了对定量研究的兴趣。

我所就读的美国加州大学洛杉矶分校社会学系，在定量研究方面有着系统的博士训练课程。不论研究兴趣是定量还是定性的，所有的研究生第一年的头两个学期必须修两门中级统计课，最后一个学期的系列课程则是简单介绍线性回归以外的其他统计方法，是选修课。希望进一步学习定量研

究方法的可以在第二年修读另外一个三学期的系列课程，其中头两门课叫“调查数据分析”，第三门叫“研究设计”。除此以外，还有如“定类数据分析”、“人口学方法与技术”、“事件史分析”、“多层线性模型”等专门课程供学生选修。该学校的统计系、心理系、教育系、经济系也有一批蜚声国际的学者，提供不同的、更加专业化的课程供学生选修。2001 年完成博士学业之后，我又受安德鲁·梅隆基金会资助，在世界定量社会科学研究的重镇密歇根大学从事两年的博士后研究，其间旁听谢宇教授为博士生讲授的统计课程，并参与该校社会研究院(Institute for Social Research)定量社会研究方法项目的一些讨论会，受益良多。

2003 年，我赴港工作，在香港科技大学社会科学部，教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课“社会科学里的统计学”(Statistics for Social Science)为所有研究型硕士生和博士生的必修课，而第二门课“社会科学中的定量分析”为博士生的必修课(事实上，大部分硕士生在修完第一门课后都会继续选修第二门课)。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的“消费者”而非“生产者”。作为“消费者”，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的

方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少量重复，但各有侧重。“社会科学里的统计学”（Statistics for Social Science）从介绍最基本的社会研究方法论和统计学原理开始，到多元线性回归模型结束，内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型，以及线性回归模型的假设和模型诊断。“社会科学中的定量分析”则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法，将重点放在因变量为定类数据的分析模型上，包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型，包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生，一直鼓励和支持我将两门课的讲稿结集出版，并帮助我将原来的英文课程讲稿译成了中文。但是，由于种种原因，这两本书拖了四年多还没有完成。世界著名的出版社 SAGE 的“定量社会科学研究”丛书闻名遐迩，每本书都写得通俗易懂。中山大学马骏教授向格致出版社何元龙社长推荐了这套书，当格致出版社向我提出从这套丛书中精选一批翻译，以飨中文读者时，我非常支持这个想法，因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种

语言的精准把握能力,还要有对实质内容有较深的理解能力,而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容,只有语言能力是远远不能胜任的。在短短的一年时间里,我们组织了来自中国内地及港台地区的二十几位研究生参与了这项工程,他们目前大部分是香港科技大学的硕士和博士研究生,受过严格的社会科学统计方法的训练,也有来自美国等地对定量研究感兴趣的博士研究生。他们是:

香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛,硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊,应用社会经济研究中心研究员李俊秀;香港大学教育学院博士研究生洪岩璧;北京大学社会学系博士研究生李丁、赵亮员;中国人民大学人口学系讲师巫锡炜;中国台湾"中央"研究院社会学所助理研究员林宗弘;南京师范大学心理学系副教授陈陈;美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛;美国加州大学洛杉矶分校社会学系博士研究生宋曦。

关于每一位译者的学术背景,书中相关部分都有简单的介绍。尽管每本书因本身内容和译者的行文风格有所差异,校对也未免挂一漏万,术语的标准译法方面还有很大的改进空间,但所有的参与者都做了最大的努力,在繁忙的学习和研究之余,在不到一年的时间内,完成了三十五本书、超过百万字的翻译任务。李骏、叶华、张卓妮、贺光烨、宋曦、於嘉、郑冰岛和林宗弘除了承担自己的翻译任务之外,还在初稿校对方面付出了大量的劳动。香港科技大学霍英东南沙研究院的工作人员曾东林,协助我通读了全稿,在此

我也致以诚挚的谢意。有些作者，如香港科技大学黄善国教授、美国约翰·霍普金斯大学郝令昕教授，也参与了审校工作。

我们希望本丛书的出版，能为建设国内社会科学定量研究的扎实学风作出一点贡献。

吴晓刚

于香港九龙清水湾

# 序

当乔治·盖洛普(George Gallup)正确预测到富兰克林·罗斯福(Franklin D. Rooselvelt)成为1936年美国总统选举赢家时,公众民意调查进入了科学抽样的时代,那时使用的方法是配额抽样,一种代表目标总体的非概率抽样方法。但是,同样的方法却错误地预测托马斯·杜威(Thomas Dewey)将成为1948年的总统,但实际上哈利·杜鲁门(Harry S. Truman)才是最终的赢家。这种方法之所以失败,是因为配额抽样是非概率的,而且盖洛普的配额抽样框基于1940年的人口普查,忽略了二次大战期间的城市移民。

今天的调查抽样方法和早些时候相比已有了很大进步,现在我们依赖于复杂的概率抽样设计。一个关键的特征就是分层:目标总体被分成一系列不同阶层的子总体,阶层内的样本量由抽样者控制且通常与该阶层的人口规模成比例。另一个特征是整群和多阶段抽样:先抽取组群——即在不同阶层选到的一系列不同等级的集群,到最后才在这些“最后阶层”的集群内部选取个体成员。比如,美国综合社会调查使用的就是分层多阶段集群抽样设计。Kalton(1983)为调查

抽样提供了很好的入门介绍。

当调查设计具有这个复杂的性质时,数据的统计分析就不再是一个简单的运行回归(或任何其他的模型)分析的事情了。现在的调查全都伴随着抽样权重以支持正确的统计推断,大部分关于统计分析的教材,通过假定简单随机抽样而没有处理抽样权重的问题,而这个被忽略的抽样权重可能对统计推断产生重要的影响。在最近的二三十年间,数据分析的统计方法也已取得了巨大进展。这些很可能就是 Michael S. lewis-Beck 选择为《分析复杂调查数据》(*Analyzing Complex Survey Data*)出第二版的原因。

Lee 和 Forthofer 这本书的第二版为我们在调查抽样设计和调查数据分析的联合方面提供了最新的情况。作者在本书开头先回顾了调查抽样设计常见的类型,并通过解释什么是抽样权重及其如何产生和调整,揭开了抽样权重的神秘面纱。然后他们详细讨论了方差估计以及考虑抽样权重后的复杂截面调查数据的初级和多变量分析等主要问题。他们重点关注了基于设计的方法,这种方法直接在分析中涉及抽样设计(虽然他们也讨论了基于模型的视角,该视角在某些分析中能扩充基于设计的方法),他们还以流行的软件为例阐释了这种方法的使用。读者将会在以样本为基础作出统计推断的实践中发现本书的巨大益处。

廖福挺

# 目 录

# 第1章

# 概　论

调查分析似乎通常是在所有的样本观察值都以同等的概率被独立选中的情况下实行的。如果在数据收集中使用的是简单随机抽样(SRS),这种分析就是对的,但实际上样本选择比SRS复杂得多。某些样本观察值可能比其他观察值以更高的概率被选中,某些之所以被包括在样本中,是因为他们属于某个特定组别的成员(比如家庭),而不是被独立选取的。我们可以在调查数据的分析中简单地忽略这些与SRS相悖的事实吗?使用调查数据分析统计书中的标准技术是否合适?或者是否有特别的方法和计算机程序更适合复杂调查数据的分析?这些问题将在随后的章节中进行论述。

今天典型的社会调查反映了统计理论和关于社会现象的知识两者间的结合,过去70年间从许多不同的调查中获得的经验塑造了它的发展。社会调查的进行是为了满足用以讨论社会、政治和公共卫生议题的信息需要。为了满足这种信息需要,在政府内部和外部都设立了调查机构。但是,在早期提供这些信息的尝试中,调查小组们最关心的是实地调查中的操作问题,如抽样框的建立,员工培训/监督,以及成本的降低等,理论上的抽样议题获得的只是次级重视(Ste-

phan, 1948)。随着这些实际问题得到解决,现代抽样实践方法的发展远远超越了 SRS。复杂抽样设计已经走在了前面,随之而来的是一系列分析问题。

因为早期调查通常需要的只是描述性统计,所以很少有人对分析问题感兴趣。更近一些时期,社会和政策科学家对分析研究的需要已经增加,那些并不参与数据收集过程的研究人员,正利用可用的社会调查数据对各种不同的现实议题进行分析。这个传统上被称为二手资料分析(Kendall & Lazarsfeld, 1950)。研究者通常并没有对复杂抽样设计的发展给予应有的关注,并假设这些设计对将要使用的分析程序没有什么影响。

二手资料分析中统计技术使用的增加以及近年来对数线性模型、logistic 回归和其他多变量技术的使用(Aldrich & Nelson, 1984; Goodman, 1972; Swafford, 1980),并没有把设计和分析更紧密地结合起来。这些技术断定数据收集使用的是简单有放回随机抽样(SRSWR),但这个假设在应用了观测单位的分层和整群方法以及不等概率选择法的社会调查中几乎无法得到满足。因此,利用 SRSWR 假设的社会调查分析可能导致有偏差和误导性的结果。比如,Kiecolt 和 Nathan(1985)在他们关于二手资料分析的书中承认了这个问题,但是他们几乎没有就如何把抽样权重和其他设计特征融入分析之中提供什么建议。最近一篇关于公共健康和流行病学的文献研究表明:对基于设计的调查分析方法的使用正在逐渐增加,但依然处于较低水平(Levy & Stolte, 2000)。

任何对抽样进行了限制且那些限制超越了 SRSWR 所做限制的调查在设计上都很复杂,并需要特别的分析性考虑。

本书回顾了由复杂抽样调查引起的分析问题,为分析策略提供了入门介绍,并利用某些可用的软件进行了实例阐释。我们讨论的重点在于使用抽样权重以校正不同的代表性,以及抽样设计对抽样方差估计的影响,我们也对权重的产生和调整程序做了一些讨论。但许多其他重要的处理非抽样误差和缺失数据的议题并没有在本书中得到充分说明。

本书介绍的最基本的方法是分析复杂调查数据的传统方法。这种方法现在被称为基于设计的(或基于随机化的)分析。另一种不同的分析复杂调查数据的方法,是所谓的基于模型的分析。正如统计学的其他领域,近些年来基于模型的统计推断在调查数据分析中已引起了更多的注意。这些构造模型的方法在调查数据分析的不同步骤——定义参数、定义估计值和估计方差——中得以介绍;但是,并没有普遍被接受的模型选择或使某个特定模型有效化的规则。

然而,对基于模型的方法的理解对调查数据分析者扩充基于设计的方法来说非常重要。在某些情况下,这两种方法会产生相同的结果,但在其他情况下结果却不同。基于模型的方法可能在描述性数据中没什么作用,但在推断性分析中却是有用的。我们将在适当的地方介绍基于模型的视角,并进一步提供与这些问题的处理方式相关的参考资料。恰当地实施基于模型的分析需要对一般统计模型知识的掌握,而且还需要从调查统计员那里得到某些参考信息。此书中与这种不同的方法或相关题目有关的章节用星号(*)标出。

自本书第一版出版以来,对复杂调查数据进行分析的软件情况已有相当大的改进。方便用户的程序现在很容易就可获得,许多常用的统计方法现在也已合并成不同的程序

包,包括 logistic 回归和生存分析。这些程序将在这个版本中用实例加以介绍。这些程序可能比其他标准软件更容易被错误地使用。本书所讨论的题目和议题将为避免调查数据分析中的陷阱提出一些指导原则。

在我们的叙述中,我们假定读者在一定程度上熟悉诸如简单随机抽样、系统抽样、分层随机抽样和简单二阶段整群抽样等抽样设计方法。对这些设计方法详细的描述可参见 Kalton(1983)和 Lohr(1999)的著作。我们还假定读者对标准统计方法以及某种标准统计程序包如 SAS 或 STATA 有大致的了解。

# 第2章

# 抽样设计和调查数据

我们对调查数据的分析重点关注满足两个基本要求的样本设计。首先,我们只关心概率抽样,其中总体里面的每个成员均有已知的不等于0的概率被选中在样本里面。这是对某个给定的设计把统计理论应用到调查估计量的特性的推导中的基础。其次,如果某个样本要从某个总体中被抽取出来,就有必要建立一个罗列了合适抽样单位的抽样框,且这些抽样单位必须涵盖总体中的所有成员。如果罗列所有总体成员是不可行或不现实的,那么某些总体成员的集群也可用做抽样单位。比如,建立一个包括美国所有家庭的清单是不切实际的,但是我们可以在几个不同的阶段选择样本。第一个阶段,随机选取郡;第二个阶段,在每个被选中的郡内抽取普查区;第三个阶段,在每个被选中的普查区内选取街区。然后在最后的选择阶段,只需要对被选中的街区罗列家庭清单即可。这个多阶段设计保证了总体成员都有已知的且不等于0的被选中概率。

# 第 1 节 | 抽样方法的种类

最简单的抽样设计是简单随机抽样，它要求总体中的每个成员都有相同的被包括在样本里面的概率，且罗列了所有总体成员的清单是可以获得的。样本成员的选择可以通过有放回或无放回两种方式进行。简单有放回随机抽样特别让人感兴趣，因为它通过放回处理排除了被选中的成员间的相关性(协方差)，因而简化了统计推断。但在这个方案中，一个成员可能在样本中出现不止一次。在实际中，简单随机抽样通过无放回的方式进行，即简单无放回随机抽样(SRSWOR)，因为没有必要不止一次地从同一个成员中搜集信息。此外，SRSWOR 相比 SRSWR 给出的抽样方差更小。但是，对于只从总体成员中选取一小部分的大型调查来说，这两种方法实际上是一样的。我们将在本书中使用 SRS 一词表示 SRSWOR，除非另有说明。

SRS 设计被进一步调整以适应其他理论或实际的考虑。常用的实际设计方案包括系统抽样、分层随机抽样、多阶段集群抽样、PPS 抽样(按规模大小成比例概率抽样)和其他受控制的选择程序。这些更实际的设计在两个重要方面偏离了 SRS。首先，总体成员被选中的概率(以及由个体成员组成的不同集群的联合被选中概率)可能并不相同。其次，抽

样单位可能不同于我们感兴趣的总体成员。这些偏离使常用的估计方法和方差计算方法复杂化,而且如果没有使用恰当的分析方法,可能会导致估计和统计检验出现偏差。我们将用几个具体的抽样设计详细考虑这些偏离情况,并检验它们对调查分析的影响。

系统抽样是替代 SRS 的一种比较常用的方法,其原因在于它的简单性。它在一个随机起始点(在 1 和 $k$ 之间)之后,选择每个处于第 $k$ 个位置的成员。其程序上的任务很简单,而且我们很容易对该程序进行检查,但是要通过检验结果来核实 SRS 却很困难。它常用于多阶段抽样的最后一个阶段,如当实际调查者被指示要从某个街区的住宅清单中选择预定比例的(住宅)单位时。系统抽样程序对总体中的每个成员指定相同的被选中的概率,这确保当总体中的成员数量 $N$ 等于样本中成员数量 $n$ 的 $k$ 倍时,样本均值将是总体均值的无偏估计。如果 $N$ 并不严格等于 $nk$,就无法确保相同的概率,虽然这个问题在 $N$ 很大时可以被忽略。在这种情况下,我们可以使用循环系统抽样方案。在这个方案中,随机起始点是在 1 至 $N$ 之间选择的(任何成员都可以成为起始点),然后假定这个抽样框是循环的(列表的末端与列表的开头是相连的),选择每个处于第 $k$ 个位置的成员。但是,当抽样框中的成员是以与调查变量相关的循环方式罗列,且抽样间距正好与清单的循环周期相吻合时,系统抽样就可能导致不真实的估计。比如,如果我们对到某个诊所的每隔 40 个病人中抽取 1 个,而平均每天的病人流量几乎就是 40,那么得到的系统抽样将只包括那些在一天中的某个特定时间到该诊所的人。这样的样本可能无法代表这个诊所的所有病人。

再者，即使当清单是随机排序时，与SRS不同，总体成员的不同的集合也可能具有不同的被选中的概率。比如，在系统抽样中，同时包含第$i$个和第$i+k$个成员在样本中的概率是$1/k$，但同时包含第$i$个和$i+k+1$个成员在样本中的概率为0。这使方差计算复杂化了。另一个观察系统抽样的方式是，相当于从$k$个有规则地形成的包含$n$个成员的集群中选取一个。(不同集群之间的)抽样方差无法从选到的这个集群中估计出来。因此，系统抽样中的方差估计需要特殊的策略。

用于克服这些系统抽样问题的一种修正方法是重复系统抽样(Levy & Lemeshow, 1999:101—110)。这种方法不是只从清单中进行一次抽样来选取一个系统的样本，而是分几次选取几个更小的系统性的样本，而每一次选取都以不同的起始点从头到尾进行选择。这个程序不仅避免了可能出现的抽样框的周期性问题，而且允许直接从数据中进行方差估计。从所有子样本中得到的估计值的方差可以通过从每个子样本中各自的不同估计值的变异情况中估计出来。重复系统抽样的方法为复杂调查的方差估计提供了一种策略，这将在第4章进一步讨论。

分层随机抽样把总体成员划成不同的阶层，每个阶层里又有不同的样本。它的使用有几个原因：(1)如果阶层是内部同质的，则该方法可以减少抽样方差；(2)对不同的阶层可以获得不同的估计值；(3)利用阶层，实地调查的执行可以组织起来；(4)在不同的阶层内部，不同的抽样需要可以得到满足。

当不同阶层间抽样的比例是统一的时候，样本在不同阶

层间的分配是成比例的;或者,当某个更高的抽样比例应用到某个更小的阶层,从而为比较研究选取足够的对象时,样本分配是不成比例的。总的来说,分层随机抽样的估计过程比 SRS 更加复杂。通常被称为两步处理法。第一步是分别在每个阶层内部进行统计量——比如均值及其方差——的计算。然后以反映每个阶层内总体比例的权重为基础把这些估计值结合起来。正如我们后面将要讨论的,这种方法也可称为利用加权统计量的一步处理法。在成比例分层抽样的情况下,估计过程可以简化,但在方差估计时必须把阶层考虑进来。

划分阶层需要在抽样框中找到用来分层的变量的信息。当这些信息找不到时,分层方法就不能应用到抽样设计中。但可以在数据搜集完成之后进行分层以改善估计的精确度。这种事后分层方法的使用,通过把样本的人口构成调整到可知的总体构成水平,从而使样本更具总体代表性。通常,为了利用人口普查数据,在事后分层中常使用诸如年龄、性别、种族和教育等人口统计学的变量。这种调整需要使用权重和不同的方差估计方法,因为在事后分层设计中,阶层的样本规模是一个随机变量(在数据收集完之后才决定的)。

集群抽样是一种实用的调查方法,因为它以由不同的成员构成的组(集群)为单位,而不是直接对个体成员进行抽样。它简化了构造抽样框的任务,降低了调查成本。正如前面所讨论的,通常使用的是一系列不同阶层的地理集群。在多阶段集群抽样中,除了在最后的抽样阶段,其他阶段中的抽样单位是由成员构成的组别。当成员的数量在不同的集群间相同的时候,估计过程也与 SRS 相同。但集群规模不同

的简单随机抽样会导致处于更小集群中的成员比处于更大集群中的成员更有可能被选中。此外,集群通常被分层以实现某些调查目的和实地的调查步骤,比如,对特殊少数人口集群的过取样。不成比例分层和不同规模集群使估计过程变得复杂。

在不同规模集群的一步集群抽样中选取自我加权的成员样本,其中一个方法是以与集群规模成比例的概率选取集群(PPS抽样)。但是,这需要知道集群的真实规模。因为真实规模通常在调查时并不可知,所以选择概率反而是与估计到的规模成比例的(PPES抽样)。比如,在某个以医院作为集群的出院调查中,病床的数量可以用来测量医院规模。PPES抽样的一个重要的后果是预期样本量将因初级抽样单位(PSU)的不同而不同。也就是说,样本量不是固定的,而是随着样本的改变而改变的。所以,样本量——样本均值计算中的分母——是一个随机变量,因此样本均值成为两个随机变量的比率。这种类型的变量,比率变量,需要特别的策略进行方差估计。

# 第2节 | 调查数据的属性

如果我们要从样本推断总体,样本选择过程就是推断过程中一个不可或缺的部分,且调查数据必须包含选择程序各重要层面的信息。考虑到多数社会调查对SRS的偏离情况,我们不仅要把调查数据看做测量的记录,还要注意它具有不同的代表性和结构安排。

抽样权重用于反映样本成员不同的被选择概率。抽样权重的产生需要密切关注每个抽样阶段和每个阶层各自的被选择概率。此外,它还涉及在样本的组别内校正不同的应答率,并通过人口统计学的变量把样本分布调整到已知的总体分布水平,即事后分层调整。还有,不同的分析单位可能需要不同的抽样权重。比如,在某个社区调查中,可能需要对个人层次的分析创建个人权重,对家庭数据的分析创建家庭权重。

当以下某种自加权设计被使用时,我们也可能会对未加入权重比较放心。一步集群抽样中正确的PPS抽样会产生一个自加权成员样本,就像在SRS设计里面一样。自加权也可以在两步设计中实现,即当在第一阶段中使用了正确的PPS抽样,且在每个被选中的PSU中选取固定数量的成员时。如果简单随机抽样在第一阶段中使用,且在第二阶段选

取固定比例的成员，也会出现相同的结果(Kalton, 1983:第5章和第6章)。但在实际中，自加权特征会遭到无应答和抽样框中的可能误差的破坏。这种非计划中的自加权可能产生偏差，但这种偏差却不大可能通过对样本数据进行检验来评估。事后分层和无应答调整是用于减少这种误差的两种方法。事后分层涉及权重的分配，以把不同子群中的样本比例调整为与总体比例相一致。无应答调整对那些参与了调查的人提高了权重，以弥补那些具相似特征但没有应答的人。因为要通过加权而实现无应答和事后分层调整，所以即使当自加权设计被应用时权重的使用仍几乎不可避免。

抽样设计影响了标准误差的估计，因此也必须结合到分析中。仔细检查那些我们熟悉的、在统计教科书中出现且内置于大部分计算机程序包中的标准误差计算方程，我们就能发现它们都是基于 SRSWR 设计的。这些方程相对比较简单，因为在假设了成员的选择具独立性之后，成员间的协方差为0。但对这些方程如何加以修正以调整其他复杂的抽样设计却并不清楚。

为了更好地理解对方差方程进行调整的需要，我们对几个不同的抽样设计检验一下方差方程。首先考虑从 SRSWOR 设计中得到的样本均值的方差。那个我们熟悉的样本均值 $\bar{y}$ 的方差方程(通过 SRSWR 从有 $N$ 个成员、均值为 $\bar{Y}$ 的总体中抽取一个包括 $n$ 个成员的样本)在初级统计教科书中等于 $\sigma^2/n$，其中 $\sigma^2 = \sum (Y_i - \bar{Y})^2/N$。这个方程在 SRSWOR 设计中需要被修改，因为这里一个成员的选择已经不再独立于另一个成员的选择。由于不允许重复选择，在第 $i$ 个和第 $j$ 个样本成员间存在一个负的协方差 $[-\sigma^2/(N-1)]$。合并

$n(n-1)$ 乘以协方差，SRSWOR 样本均值的方差就变成了 $\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$，这个比 SRSWR 中的方差小，是其 $(N-n)/(N-1)$ 倍。把 $\sigma^2$ 的无偏估计值用 $[(N-1)s^2/N]$ 代替，SRSWOR 样本均值的方差估计值就变成：

$$\hat{V}(\bar{y}) = \frac{s^2}{n}(1-f) \qquad [2.1]$$

其中，$s^2 = \sum(x_i - \bar{x})^2/(n-1)$，$f = n/N$。$(N-n)/(N-1)$ 和 $(1-f)$ 两者都称为有限总体校正(FPC)因素。在一个大总体中，协方差会非常小，因为抽样部分非常小。因此 SRSWR 和 SRSWOR 设计会产生几乎相同的方差，这两种程序在所有实际目标中可以被看做是相同的。

分层抽样通常被认为是一个更有效的设计，因为如果使用恰当，它给出的方差比类似的 SRS 给出的方差更小。因为阶层间的协方差为 0，所以样本估计值的方差是从阶层内的变异情况——以阶层的样本量和阶层权重为基础结合起来——推导出来的。分层样本方差的取值取决于阶层样本量的分布。一个最优(或者 Neyman)分配产生的抽样方差要小于或等于基于 SRS 而产生的方差，但在极端少数的情况下也有例外。对其他不成比例的分配来说，当阶层内的有限总体校正因素不能被忽视时，抽样方差可能比那些基于 SRS 的方差要大。因此，相比 SRS，分层的方法并不一定总会降低抽样方差。

集群抽样设计常常导致比从 SRS 中得到的方差更大的抽样方差。这是因为在自然形成的集群内部其成员通常比较相似，因此在集群内各成员间产生一个正的协相关。集群

内的同质性用组内相关系数(ICC)——集群内所有可能的成员配对组合间的相关性——测量。如果集群是随机形成的(即如果每个集群都是成员的随机样本),ICC 就等于 0。在许多自然集群中,ICC 是正的,因此抽样方差会比从 SRS 设计中得到的更大。

在一个复杂的设计中,很难就抽样方差的相对大小进行归纳,因为必须对分层和集群的联合效应以及抽样权重的效应进行评估。所以,调查数据中所有的观察值都必须看做是一个特定的包含抽样权重和结构安排的抽样设计的产物。除了抽样权重以外,阶层和集群的标识(至少 PSU)也应该包含在样本调查数据中。这些要求的原因我们也会在下文中详细论述。

复杂调查中方差计算的一个复杂情况来源于权重的使用。因为任何加权后方差估计之分母中的权重总和并不固定,反而随样本的不同而改变,所以该估计就变成了两个随机变量的比率。一般来说,比率估计值是有偏差的,但如果权重的变化相对较小或样本量比较大,这个偏差就可以忽略不计(Cochran, 1977:第 6 章)。因此,在大型社会调查中,比率估计的有偏差问题并不严重。但是,因为这个偏差,使用均方差——方差和偏差平方的和——比使用方差更合适。但因为这个偏差通常是可以忽略的,因此在本书中我们使用“方差”一词,尽管其所指实为均方差。

# 第 3 节 | 调查数据的另一种不同看法*

迄今为止,对调查数据属性的描述是从基于设计的视角展开的——样本数据是利用某个特定的样本选择设计方案从一个有限的总体中抽取的观察值。抽样设计明确了每个潜在样本的选择概率,然后某个合适的估计方法被选中用以反映这种设计。如概论中所述,基于模型的视角提供了样本调查数据的另一种看法。有限总体中的观察值被看做是从某个模型中产生的随机变量(服从某种概率分布的随机变量)的现实对应值。这假定的概率模型连接了样本中的单位和不在样本中的单位。在基于模型的方法中,样本数据用于预测无法观测到的值,因此推论可以被认为是预测问题(Royall, 1970、1973)。

这两种不同的视角在可以合理地假设样本观测值是在一个均值为 $\mu$、方差为 $\sigma$ 的正态分布中独立同分布的 SRS 中可能没什么差别,但从模型的角度看,全及总体是样本中的观测值以及没有包含在样本中的观测值的总和,也就是 $Y = \sum_{i \in S} y_i + \sum_{i \notin S} y_i$。基于同均值的假设,全及总体的估计值可以写成 $\hat{Y} = n\bar{y} + (N-n)\bar{y} = N\bar{y}$,其中 $\bar{y}$ 是在该模型下未观测到的观察值的最优无偏估计值。结果它与基于设计的方法中的扩展估计值相同,即 $\hat{Y} = (N/n)\sum_{i=1}^{n} y_i = N\bar{y}$,其

中，$(N/n)$是抽样权重（SRS 中选择概率的倒数）。这两种方法都得出了同样的方差估计值（Lohr，1999：第 2.8 节）。

但是，如果采用的是不同的模型，方差估计值就可能不同。比如，在 SRS 比率[1]和回归估计的情况下，假定的模型是 $Y_i = \beta x_i + \varepsilon_i$，其中 $Y_i$ 代表一个随机变量，$x_i$ 是一个辅助变量，对这个辅助变量来说全及总体是已知的。在这个模型下，全及总体的线性估计为 $\hat{Y} = \sum_{i\in S} y_i + \sum_{i\notin S} y_i = n\bar{y} + \hat{\beta}\sum_{i\notin S} x_i$。其中，第一部分是从样本中来的，第二部分是基于假定的模型对没有观察到的单位作出的估计。如果我们把 $\hat{\beta}$ 取值为样本比率 $\bar{y}/\bar{x}$，那么就有 $\hat{Y} = n\bar{y} + \frac{\bar{y}}{\bar{x}}\sum_{i\notin S} x_i = \frac{\bar{y}}{\bar{x}}(n\bar{x} + \sum_{i\notin S} x_i) = \frac{\bar{y}}{\bar{x}}X$，其中 $X$ 是 $x_i$ 的全及总体。这就是 $Y$ 的比率估计。如果我们把 $\hat{\beta}$ 取值为估计到的回归系数，那么我们就得到了一个回归估计。虽然从基于设计的角度看比率估计被认为是轻微有偏的，但从基于模型的推理角度看如果模型正确它就是无偏的。

但是，从基于模型的方法中得到的方差估计值与从基于设计的方法中得到的稍微有所不同。基于设计的预计全及总体的方差估计值为 $\hat{V}_D(\hat{Y}) = \left(1-\frac{n}{N}\right)\left(\frac{N^2}{n}\right)\cdot\frac{\sum[y_i-(\bar{y}/\bar{x})x_i]^2}{n-1}$。基于模型的估计值为 $\hat{V}_M(\hat{Y}) = \left(1-\frac{x}{X}\right)\left(\frac{X^2}{x}\right)\frac{\sum[\{y_i-(\bar{y}/\bar{x})\}/\sqrt{x_i}]^2}{n-1}$，其中 $x$ 是全及样本，$X$ 是辅助变量的全及总体（Lohr，1999：第 3.4 节）。

当 $y_i$ 和 $x_i$ 之间的关系是过原点直线的，且 $y_i$ 关于这条

线的方差是与 $x_i$ 成比例时,比率估计模型是有效的。当 $y_i$ 和 $x_i$ 之间的相关性小于 $x_i$ 的变异系数与 $y_i$ 变异系数的比率的一半时,比率估计被认为差于(没有辅助变量的)扩展估计(Cochran, 1977:第 6 章)。因此,在调查分析中使用比率估计要求对模型假设进行检验。在实际中,当数据库包括大批变量时,比率估计会变得笨重,要为不同的估计选择不同的辅助变量。

要把基于模型的方法应用到实际问题中,我们必须能够建立一个适当的模型。如果模型是错的,那么基于模型的估计就是有偏的。在抽样中使用基于模型的推断时,我们需要通过详细分析数据来检验模型的假设。在许多情况下,检验假设是困难的。模型的恰当性在某种程度上说是判断力的问题,而且适合某个分析的模型,可能对另一个分析或另一个调查并不合适。

# 第3章

# 分析调查数据的复杂性

调查数据分析的两个根本方面是,调整样本观察值的不同的代表性和评估由样本设计选择的复杂性带来的精确度的得失。这一章介绍权重的概念并讨论样本设计的选择对方差估计的影响。为了阐释权重在调查分析中的多用性,我们列举了两个创建和调整抽样权重的实例。

# 第1节 | 调整不同的代表性：权重

有两种类型的权重在调查数据的分析中经常遇到：(1)扩展权重，它是选择概率的倒数；(2)相对权重，通过按比例缩小扩展权重以反映样本规模而获得。这一部分我们将为几种不同的抽样设计详细介绍这两种不同的权重。

考虑如下 SRS 情况：对一个清单中某总体的 $N = 4000$ 个成员用 1 到 4000 进行排号。用一张随机数字表从这个总体中抽取固定数量的成员（比如 $n = 200$），不允许重复选择（无放回）。选择概率或抽样比例为 $f = n/N = 0.5$。扩展权重是选择概率的倒数，$w_i = 1/f = N/n = 20(i = 1, \cdots, n)$，表示由一个样本观察值所代表的总体中的成员数量。这 $n$ 个被选中的成员的权重之和等于 $N$。基于样本成员，变量 $Y$ 的全及总体的估计值为：

$$\hat{Y} = \sum w_i y_i = (N/n) \sum y_i = N\bar{y} \qquad [3.1]$$

方程 3.1 表现了扩展权重在样本观察值的加权总和中的使用。由于对 SRS 中的每个成员权重是一样的，因此这个估计值可以简化为样本均值的 $N$ 倍（方程 3.1 中的最后一项）。同样，总体均值的估计值等于 $\hat{\bar{Y}} = \sum w_i y_i / \sum w_i$，即加权后的样本均值。在 SRS 中，它可以简化为

$(N/n)\sum y_i/N=\bar{y}$，表示样本均值是总体均值的估计值。但是，即使在不等概率设计中，各成员的权重不相同，全及总体和总体均值的估计值仍然是加权总和和加权平均数。

虽然扩展权重看起来适合全及总体的估计，但它可能会给样本均值和其他统计测量带来重大影响。比如，在列联表中用扩展权重的总和代替以样本规模为基础的相对频数可能会夸大数据中的置信度。要处理这个问题，可以把扩展权重按比例缩小以生成相对权重$(rw)_i$，即用扩展权重除以扩展权重的均值，$w_i/\bar{w}$。其中，$\bar{w}=\sum w_i/n$。样本中每个成员的相对权重之和等于 $n$。对 SRS 设计来说，每个成员的$(rw)_i=1$。用相对权重进行加权后的全及总体的估计为：

$$\hat{Y}=\bar{w}\sum(rw)_i y_i=(N/n)\sum y_i=N\bar{y} \qquad [3.2]$$

注意，在方程 3.2 中，相对权重被乘以扩展权重的平均数，对 SRS 得出的简化估计值与方程 3.1 一模一样。因此，在对全及总体进行估计时，扩展权重的使用比相对权重更简单。相对权重在分析性研究中更合适，但却不适用于估计总体和计算有限总体校正值。

大部分政府部门和调查机构提供的公开调查数据使用的都是扩展权重，很容易转化为相对权重。但这种转化并不是必须的，因为许多便于用户使用的调查分析统计程序会在适当的时候自动对其进行转化。

我们来分析一下分层随机抽样设计中扩展权重的使用。在这种设计中，具有 $N$ 个成员的总体基于某个变量被分成 $L$ 个阶层，每个阶层分别包含 $N_1$，$N_2$，…，$N_L$ 个成员。从这些阶层中，分别有 $n_h$ 个($h=1$，2，…，$L$)成员从第 $h$ 个阶层

中被抽取出来。当每个阶层的抽样比例相同的时候，这个分层抽样设计便具有自加权的特征。如果一共有 200 个成员从两个阶层（$N_1 = 600$ 和 $N_2 = 3400$）中按比例被抽取出来，那么抽样比例 $f = 200/4000 = 0.05$。按这个比例选取（从每个阶层中各选取 5%的样本）的结果会产生两个样本 $n_1 = 30$ 和 $n_2 = 170$，因为 $f_1 = 30/600 = 0.5$，$f_2 = 170/3400 = 0.5$。因此加权的方案与 SRS 设计中的结果完全相同。

在不成比例分层抽样设计中情况则稍微不同。比如，如果这 200 个样本成员在两个阶层中平摊，那么 $f_1$（$=100/600$）和 $f_2$（$=100/3400$）的取值就不相同，这两个阶层的成员其扩展权重也不相同，分别为 $w_{1i} = 6$ 和 $w_{2i} = 34$。扩展权重之和在第一个阶层内为 600，在第二个阶层内为 3400，总和 4000 等于总体规模。扩展权重的均值 $\bar{w} = (100 \times 6 + 100 \times 34)/200 = 20$，因此相对权重之和在第一个阶层内为 30，在第二个阶层内为 170，两个阶层总和等于样本规模。值得注意的是，这两种权重都相当于用总体在各阶层的分布状况对阶层均值（$\bar{y}_h$）进行加权（即 $\sum (N_h/N)\bar{y}_h$，标准程序）。在阶层 1 内，扩展权重和相对权重两者的总和都分别等于它们各自的全部总和的 15%，而且第一阶层包含的成员的数量也等于所有总体成员数量的 15%。

虽然我们在介绍抽样权重时用的是 SRS 和分层抽样设计，但同样的概念可以很容易扩展到更加复杂的设计的使用中。总的来说，抽样权重是选择概率的倒数，虽然我们经常会通过事后分层和无应答调整的方法对它做进一步的调整。对每个样本成员赋予抽样权重有助于对所有的抽样设计建立一个综合的估计程序。一个普遍的规则是，在调查数据分

析中所有的估计都以加了权的统计量的形式进行。除了估计总体和计算有限总体校正值,这些权重的规模在估计参数和标准误差时并不重要。

接下来,我们举两个创建/调整抽样权重的例子。第一个例子告诉我们如何通过事后分层的方法调整抽样权重,以使样本结构与总体结构一致。同样的方法也可用于调整各不同人口子群中的不同的应答率。当涉及的人口特征变量只有少数几个的时候,事后分层的方法效果较好。第二个例子告诉我们在追踪调查中,我们可以通过使用 logistic 回归模型,对许多变量的不同的损耗率进行调整。

# 第2节 | 用事后分层的方法加权

为了演示抽样权重的创建过程，我们利用美国1984年的综合社会调查(GSS)进行说明。这个调查是由美国全国民意研究中心(NORC)组织的一个复杂的抽样调查，旨在从美国普通的非集体住户的成年(大于或等于18岁)人口中获得综合社会信息。该调查使用了多阶段抽样方案以在家庭层面产生一个自加权样本。然后从每一个被抽中的家庭里面随机选取一个成年人(Davis & Smith, 1985)。最终的数据总共包括1473个观察记录可供分析使用。对家庭层面的数据来说，扩展权重可以计算为等于美国所有家庭的总数量除以1473，而被选中的家庭内的扩展权重则等于该家庭内成年人的数量。这两种权重的乘积等于样本中个体成员的扩展权重。

为了分析GSS数据，我们只需要集中讨论家庭内部的权重，因为每个家庭都具有相同的被选中的概率。个体成员的相对概率可以计算为等于这个家庭内成年人的数量除以平均每个家庭内成年人的数量(后者为2852/1473 = 1.94)。这个权重反映了在维持样本规模的基础上，一个成年人被选取到样本中的概率。

为了使样本结构与总体结构一致，我们需进一步对这个权重进行调整。这样可以提高估计的准确度，同时也可把无

应答和样本选择偏差降低到只与人口构成相关的程度。[2]如表 3.1 所示,调整因素的产生是为了使样本中个体的分布状况在年龄、种族和性别等方面符合 1984 年美国人口总体的分布情况。第一列显示美国人口普查局估计的分年龄、种族和性别的人口分布状况。第二列显示在对应的人口分类中,在选中的所有家庭里面加权后的成人数量,第三列显示的是对应的分布比例。调整因素是第一列和第三列数字的比。调整后的权重等于把调整因素乘以相对权重,这样调整后的权重的分布就与人口总体的分布一致。

**表 3.1 事后分层调整因素的产生:美国 1984 年综合社会调查**

| 统计群 | 人口分布<br>(1) | 成人权重数<br>(2) | 样本分布<br>(3) | 调整因素<br>(4) |
|---|---|---|---|---|
| 白人,男性 | | | | |
| 18—24 岁 | 0.0719660 | 211 | 0.0739832 | 0.9727346 |
| 25—34 岁 | 0.1028236 | 193 | 0.0676718 | 1.5194460 |
| 35—44 岁 | 0.0708987 | 277 | 0.0795933 | 0.8907624 |
| 45—54 岁 | 0.0557924 | 135 | 0.0473352 | 1.1786660 |
| 55—64 岁 | 0.0544026 | 144 | 0.0504909 | 1.0774730 |
| 65 岁及以上 | 0.0574872 | 138 | 0.0483871 | 1.1880687 |
| 白人,女性 | | | | |
| 18—24 岁 | 0.0705058 | 198 | 0.0694250 | 1.0155668 |
| 25—34 岁 | 0.1007594 | 324 | 0.1136045 | 0.8869317 |
| 35—44 岁 | 0.0777364 | 267 | 0.0936185 | 0.8303528 |
| 45—54 岁 | 0.0582026 | 196 | 0.0682737 | 0.8469074 |
| 55—64 岁 | 0.0610057 | 186 | 0.0652174 | 0.9354210 |
| 65 岁及以上 | 0.0823047 | 216 | 0.0757363 | 1.0867272 |
| 非白人,男性 | | | | |
| 18—24 岁 | 0.0138044 | 34 | 0.0119215 | 1.1579480 |
| 25—34 岁 | 0.0172057 | 30 | 0.0105189 | 1.6356880 |
| 35—44 岁 | 0.0109779 | 30 | 0.0105189 | 1.0436290 |
| 45—54 岁 | 0.0077643 | 37 | 0.0129734 | 0.5984774 |
| 55—64 岁 | 0.0064683 | 12 | 0.0042076 | 1.5372900 |
| 65 岁及以上 | 0.0062688 | 18 | 0.0063113 | 0.9932661 |

续表

| 统计群 | 人口分布 (1) | 成人权重数 (2) | 样本分布 (3) | 调整因素 (4) |
|---|---|---|---|---|
| 非白人，女性 | | | | |
| 18—24岁 | 0.0145081 | 42 | 0.0145081 | 0.9851716 |
| 25—34岁 | 0.0196276 | 86 | 0.0301543 | 0.6509067 |
| 35—44岁 | 0.0130655 | 38 | 0.0133240 | 0.9806026 |
| 45—54岁 | 0.0094590 | 33 | 0.0115708 | 0.8174890 |
| 55—64岁 | 0.0079636 | 30 | 0.0105189 | 0.7570769 |
| 65岁及以上 | 0.0090016 | 27 | 0.0094670 | 0.9508398 |
| 总计 | 1.0000000 | 2852 | 1.0000000 | |

资料来源：U. S. Bureau of the Census, Estimates of the population of the United States, by age, sex, and race, 1980 to 1985(Current Population Reports, Series P-25, No. 985), April 1986. Noninstitutional population estimates are derived from the estimated total population of 1984(Table 1), adjusted by applying the ratio of noninstitutional to total population(Table A1)。

这些调整因素表明，如果没有进行调整，那么GSS样本中25—34岁男性的比例低于实际比例，而非白种45—54岁男性和非白种25—34岁女性的样本比例则高于实际比例。

然后将这些调整后的相对权重用于数据分析中的成年人的比例。比如，对“是否有你能够想象到的这样一类情况，在这类情况下你会同意一位男性攻击另一位陌生男性？”这个问题持肯定答案的成年人的比例。正如表3.2上半部显示的，加权后的总体比例为60.0%，比未加权时的估计值59.4%稍微大一点。对子群体的估计结果显示，加权后估计值和未加权估计值之间的差异很小。这可能主要归因于自加权的特征——反映在大部分家庭只有两个成年人这个现实上，以及在更小程度上反映在“同意攻击”与家庭内成年人数量并没有关系这个事实上。在美国国家精神卫生研究所发起的流行病学调查(Epidemiologic Catchment Area, ECA)

中,情况则有所不同。如表 3.2 所示,在这个调查中,加权后的所有障碍和焦虑性障碍流行情况的估计值分别比未加权估计值低 20%和 26%。

**表 3.2 两个调查中加权后估计和未加权估计的比较**

| 调查及变量名称 | 加权后估计 | 未加权估计 |
|---|---|---|
| 1. 美国综合社会调查(GSS) | | |
| (同意“攻击”的比例) | | |
| 总体 | 60.0 | 59.4 |
| 分性别 | | |
| 男性 | 63.5 | 63.2 |
| 女性 | 56.8 | 56.8 |
| 分教育程度 | | |
| 大学 | 68.7 | 68.6 |
| 高中 | 63.3 | 63.2 |
| 其他 | 46.8 | 45.2 |
| 2. 美国国家精神卫生研究所流行病学责任区调查(Epidemiologic Catchment Areas Survey) | | |
| (精神障碍流行率) | | |
| 所有障碍 | 14.8 | 18.5 |
| 焦虑性障碍 | 6.5 | 8.8① |

资料来源:Epidemiologic Catchment Areas Survey,数据来自 Lee、Forthofer & Lorimor(1986),表 1。

最后,应该对这些调整后的权重进行检查,看看是否存在极端大的取值。极端的离异值可能意味着某些事后分层的样本规模太小以至于不可靠。在这种情况下,某些小的事后分层可能需要合并起来,或者需要使用某些排列程序以消除那些极端值(Little & Rubin, 1987:59—60)。

① 原书中未加权估计的数据,所有障碍为 8.8,焦虑性障碍为 18.5,原书有误。——译者注

# 第3节｜在追踪调查中调整权重

追踪调查在社会科学研究中很常用。但遗憾的是，对所有的初始受访者进行持续的跟踪几乎不可能。他们中的某些人可能已经去世、搬迁，或者拒绝参与追踪调查。这些事件并不是随机分布的，在追踪调查中不同的损耗率可能导致选择偏差。基于一些人口特征变量，我们可以像在事后分层中那样采用相同的方式进行调整，但我们可充分利用初始调查中大量的潜在估计变量进行追踪调查中的损耗调整。分层的策略可能并不是很适合大量的估计变量，但是 logistic 回归模型提供了一种方式以包含若干变量。以初始调查数据为基础，我们可以建立一个 logistic 回归模型，通过使用初始调查中的一系列精心挑选过的预测变量和常用的人口特征变量来预测损耗(二分类变量)。然后，基于原始样本中受访者特征的 logit 估计值可以用于调整在追踪调查中每个受访者的初始权重。实际上，这种方法可以通过不同程度地调高那些在追踪调查中成功接触到的人的权重，来弥补那些丢失掉的受访者，以此排除选择偏差，以使之降低到只与模型中的变量相关的程度。随着模型使用更多恰当的变量，损耗调整可以比初始调查中的无应答调整更加有效。对由“追踪调查中的丢失”带来的无应答(或其他类型的无应答)进行校

正的事后分层并不能保证一定能改善估计。但是,如果“追踪调查中的丢失”与在调整程序中用到的变量相关,那么调整后的估计应该比未调整的估计更好。

我们来看某个社区精神健康调查(ECA 调查的其中一个场所)的例子。只有 74.3%的初始受访者在第一轮的年度追踪调查中成功受访。这个损耗程度太大以至于我们没法忽略不计。因此,我们挑选了一些预测变量,对损耗(1=丢失,0=受访)进行了 logistic 回归分析,结果如表 3.3 所示。卡方值说明,模型中的变量与损耗显著相关。如表 3.3 所示,效

**表 3.3 追踪调查中损耗调整的 logistic 回归模型**

| *Logistic Regression Model* | | | | *Survey-Related Information* | | |
|---|---|---|---|---|---|---|
| *Factors* | *Category* | *Variable* | *Beta Coefficient* | | | |
| Intercept | | - | -0.737* | Initial survey | | |
| Age | 18-24 yrs | AGE1 | 0.196 | Design: | | Multistage |
| Sampling | | | | | | |
| | 25-34 yrs | AGE2 | 0.052 | Sample size: | | 4967 |
| | 35-44 yrs | AGE3 | -0.338* | Weighted sum: | | 300113 |
| | 45-54 yrs | AGE4 | -0.016 | | | |
| | 55 and over | - | | Follow-up survey | | |
| Marital status: | Sep./divorced | MAR | 0.051 | Sample size: | | 3690 |
| (74.3%) | | | | | | |
| | Other | - | | Sum of attrition- | | |
| Gender: | Male | SEX | 0.084 | adjusted weights: | | 300172 |
| | Female | - | | Adjusted sum: | | 300113 |
| Race | White | RACE1 | -0.668* | | | |
| | Black | RACE2 | -0.865* | Comparison of attrition-adjusted | | |
| and Other | - | | | attrition-unadjusted | | |
| estimates | | | | | | |
| Socioeconomic | 1st quartile | SES1 | 0.534* | | | |
| status | 2nd quartile | SES2 | 0.389* | Disorders | Unadjusted[a] | Adjusted[b] |
| Diff. | | | | | | |
| | 3rd and 4th | | - | | | |
| Family size: | One | SIZE1 | 0.033 | Any disorder | 39.2% | 43.8% |
| -4.6% | | | | | | |
| | 2-4 members | SIZE2 | -0.003 | Major depre. | 15.0 | 18.1 |
| -3.1 | | | | | | |
| | 5 or more | - | | Cog. Impair. | 1.8 | 1.3 |
| 0.5 | | | | | | |
| Diagnosis: | Cog. Impair. | DX1 | 0.472* | Phobias | 8.4 | 7.6 |
| -0.4 | | | | | | |
| | Schizophrenia | DX2 | -0.049 | Alcohol abuse | 13.9 | 13.7 |
| -4.2 | | | | | | |
| | Antisocial | DX3 | 0.412* | Schizophrenia | 0.6 | 0.6 |
| -0.4 | | | | | | |
| | Anorexia | DX4 | 2.283* | | | |
| | No disorder | - | . | | | |

Likelihood ratio chi-square(16) = 198.0, $p < 0.00001$.

注:a. 基于最初的权重。

b. 基于损耗调整后的权重。

* $p<0.05$。

果编码需要每个变量都省略其中一个取值。基于估计到的beta系数，对每个受访者都计算了logit估计值（$\hat{\lambda}_i$, $i=1, 2, \cdots, n$）。然后这些值被转化成损耗的预测比例 $\hat{p}_i = 1/(1+e^{-\hat{\lambda}_i})$。通过把在追踪调查中丢失的人的权重设为0，以及把成功受访者的权重除以（$1-\hat{p}_i$），这些成功受访者的初始权重增加了。

正如表3.3所示，对保留下来的成员调整后的权重总和为300172，只比初始调查中的权重总和大59，这个数据意味着这种方法的使用很合理（如果调整后的权重总和与初始调查中的权重总和之间存在很大的差异，调整过程就可能值得担忧）。这些调整后的权重再次被调整以与初始调查达到一致。为了说明使用损耗调整权重的作用，如表3.3所示，我们对六种精神障碍的流行率分别进行了使用和不使用损耗调整权重这两种不同的估计。我们发现，使用了调整权重后，所有障碍（由DSM-III定义的障碍）的流行情况比没有调整时的情况大约高了五个百分点。

# 第4节 | 评估精确度的得失:设计效应

如前一章所述,SRSWOR样本均值的方差等于SRSWR样本均值的方差乘以有限总体校正因素$(1-f)$。因此,SRSWOR抽样方差与SRSWR抽样方差的比率等于$(1-f)$,反映了(与使用SRSWR相比)使用SRSWOR的影响。这个把某个特定抽样设计中某个统计值的方差与SRSWR中的进行比较的比率被称为那个统计值的设计效应。它被用于评估与SRSWR设计方案相比,现有使用方案中的样本估计其精确度的得失。设计效应小于1意味着若要得到与SRSWR相同的精确度,需要更少的观察值;设计效应大于1意味着要得到与SRSWR相同的精确度,需要更多的观察值。在SRSWOR设计中,设计效应小于1,但当抽样比例很小的时候接近于1。由于研究者习惯上常用SRSWOR代替SRSWR,因此他们倾向于以SRSWOR而不是SRSWR为基础来计算设计效应。此外,在复杂调查中,设计效应的计算通常基于SRSWOR设计中加权后统计值的方差。我们在后文中也将如此计算。

把设计效应这个概念与样本规模结合起来,有效的样本规模可以定义为实际样本规模除以设计效应。如果某个抽样设计的设计效应大于1,那么实际上为了统计分析样本规

模就会减少，因此导致更大的抽样误差。换言之，当设计效应大于1时，有效样本规模小于实际样本规模。

我们来检验一下更加复杂的抽样设计中的设计效应。大家已经知道分层随机抽样中抽样误差的性质和估计情况，以及在哪些条件下分层会产生比SRS更小的方差。但是，分层通常与其他的设计特征一起使用，如阶层内不同阶段的集群抽样。如前面讨论的，集群会增加抽样误差。在许多实际的抽样设计中，分层的作用会因集群的作用而减弱。可惜我们无法在理论上分别从分层和集群的特征中评估设计效应，反而必须从数字上大致地判断它们的联系效应。

接下来的例子说明了在一个相对简单的情况下设计效应的计算。考虑一个单步骤集群抽样，其中所有的集群大小规模一样。假设在一个高中学校有 $N$ 个英语班级，每个班级有 $M$ 个学生。从这 $N$ 个班级中用SRS选择 $n$ 个班级，然后要求这 $n$ 个被选中的班级里的所有学生都报告自今年年初以来他们读过的书的数量。在总体中学生的数量为 $NM$，在样本中学生的数量为 $nM$。抽样比例为 $f = nM/NM = n/N$。

因为班级大小相同，平均每个学生所读的书的数量（总体均值）等于 $N$ 个班级均值的平均数。这 $n$ 个样本班级可以看做具 $N$ 个均值的总体的一个随机样本，这个随机样本具有 $n$ 个均值。因此样本均值（$\bar{\bar{y}}$）对总体均值（$\bar{\bar{Y}}$）来说是无偏的，而它的方差，应用方程2.1等于：

$$\hat{v}(\bar{\bar{y}}) = \frac{s_b^2}{n}(1-f) \qquad [3.3]$$

其中，$s_b^2 = \sum(\bar{y}_i - \bar{\bar{y}})^2/(n-1)$，是集群均值的方差估计值。或者，方程3.3可以依据估计到的ICC（$\hat{\rho}$）写成（Coch-

ran，1977:第 9 章)：

$$\hat{v}(\bar{\bar{y}}) = \frac{s^2[1+(M-1)\hat{\rho}]}{nM}(1-f) \qquad [3.4]$$

其中 $s^2 = \sum\sum (y_{ij} - \bar{\bar{y}})^2/(nM-1)$，是样本成员的方差。如果方程 3.4 除以 SRSWOR 样本(样本规模为 $nM$)的均值方差$\left[\hat{v}(\bar{\bar{y}}) = \frac{s^2}{nM}(1-f)\text{，方程2.1}\right]$，那么这个集群抽样的设计效应为 $1+(M-1)\hat{\rho}$。

当 ICC $=0$ 时，设计效应为 1；当 ICC $>0$ 时，设计效应大于 1。如果集群的形成是随机的，那么 ICC $=0$；当每个集群内部所有成员有相同的取值时，ICC $=1$。社区调查中用到的大部分集群由相同地区的住宅构成，这样对该调查中许多变量产生的 ICC 是正数。通常来说，社会经济变量的 ICC 比人口特征变量如年龄性别等的 ICC 大。

对更加复杂的抽样设计的设计效应进行评估并不是一个用统计教科书中的方程就可以完成的例行任务；相反，它需要特殊的技术，而这些技术要使用一些不常见的策略。下一章将会介绍用于估计复杂调查中统计值的抽样方差和不同调查的设计效应的几种策略。

# 第5节 | 调查数据分析中抽样权重的使用

如上面所讨论的,抽样权重可用于计算点估计。所有的点估计都表现为加权后的统计值的形式。从上述讨论中可见,抽样权重的这种用法使得抽样权重的创建和调整显得很有道理,尤其是对描述性分析来说。但是,在分析性研究中,抽样权重的使用并不如在描述性分析中那样清楚。如第2章最后一部分所述,关于调查数据有 $n$ 种不同的视角。从基于设计的视角看,抽样权重的使用在描述性研究和分析性研究中都非常重要。推论是以对有限总体进行重复抽样为基础的,而用于推论的概率结构是由表明被包含在样本之中的随机变量界定的。但是,从基于模型的视角看,样本选择方案对于在某个具体的模型下进行推论是无关紧要的。如果总体中的观察值确实符合模型,那么只要模型中的因变量只通过模型中的自变量[①]而影响选择概率,抽样设计就不会有任何影响。许多学者已经对在哪些条件下做推论时可以忽略抽样方案的作用做了大量的研究(Nordberg, 1989; Sugden & Smith, 1984)。而且自20世纪70年代初期起,很多

① 而不是其他没有被包含在模型中的因素。——译者注

调查统计学家就已经针对这两种不同的视角进行过争论(Brewer, 1999; Brewer & Mellor, 1973; Graubard & Korn, 1996、2002; Hansen、Madow & Tepping, 1983; Korn & Graubard, 1995a; Pfeffermann, 1996; Royall, 1970; Sarndal, 1978; T. M. F. Smith, 1976、1983)。

好的调查数据的分析需要对这两种视角有一个全面的理解,并且还要考虑到某些实际问题,尤其是在社会调查中。基于模型的方法与其他统计分析领域日益增多的模型推论一致,而且确实提供了一些理论优势。基于模型的估计可以用于相对较小的样本之中,甚至可用于非概率样本中。此外,基于模型的分析可以用标准的统计软件如 SAS 和 SPSS 执行,而不需要依赖如 SUDAAN 及其他本书所提及的调查软件包。但是,基于模型的方法假定该模型正确地反映了自然的真实状态。如果模型设定是错的,那么分析就会有偏差,对数据的解释就会有误。可惜在社会调查中很难对总体中的所有观察值建立用理论推导出来的模型。此外,模型中相关变量的缺失在调查数据的二手分析中也是一个大问题,因为研究者并不能获得所有的相关变量。因此,基于模型的推论其主要的挑战是建立一个符合分析目的的正确的模型。

我们已经知道加权后的分析会受到具有极大权重的观察值的严重影响(极大权重的产生经常来自无应答和事后分层调整,而不是选择概率)。此外,加权的另一个局限在于方差的增加。通常,当权重的变异较大时方差会增加很多。如果实际上并没有减少偏差的必要,那么使用加权后的估计就会损失一些东西,相比未加权分析,加权后的分析也就会变得低效。Korn 和 Graubard(1999:第 4 章)讨论了处理总体

参数的加权和未加权估计的不同问题，并给出了一种检验加权估计的低效性的方法。他们建议，如果低效性并不是严重到不可接受，那么就用加权后分析，以避免未加权分析中的偏差；如果低效性太严重以至于不能接受，那么就用未加权分析，并用调查设计变量——包括权重——以充实模型从而减少偏差。但是，把设计变量结合到模型之中通常会出现问题，因为把设计变量作为附加的解释变量放到模型中可能会与科学的分析目标相矛盾。比如，当分析的目标是检验卫生措施与风险系数两者的相关性时，把设计变量放在模型中可能会干扰这两者的相关路径。

在复杂的大型调查中，通常不可能把所有的设计信息都包含在模型之中，尤其在对抽样权重进行了无应答和事后分层调整后（Alexander, 1987）。把设计变量放到模型中的方法还有另一个操作性的问题，那就是数据常常没有相关的信息。并不是所有与设计相关的信息研究者都能得到。大部分公开的调查数据只有初级抽样单位（PSU），而没有次级集群单位（如普查小区或者电话交换台）。出于保密性考虑，他们通常不可能提供次级抽样单位的信息。

在基于模型的分析中，我们必须避免可能的模型设定错误和相关解释变量的缺失。抽样权重的使用（基于设计的分析）可为此避免错误的模型设定（DuMouchel & Duncan, 1983; Pfeffermann & Homes, 1985）。Kott（1991）指出，在线性回归中需要使用抽样权重，因为在大部分二手分析中对调查数据解释变量的选择具有局限性。本书第 6 章最后一节将对使用抽样权重的利弊做进一步讨论。

# 第4章

# 方差估计的策略

对调查统计值进行方差估计很复杂,其原因不仅在于抽样设计的复杂性(如前面章节所述),也因为统计值的形式。即是在 SRS 设计中,某些统计值的方差估计也需要非标准化的估计技术。比如,标准教材明显没有涉及中位数的方差,而且比率估计值的抽样误差(参考注释[1])也很复杂,因为分子和分母同时都是随机变量。某些在一般的教科书中没有提及的方差估计方法就有足够的灵活性,可以同时顾及抽样设计的复杂性和统计值的各种不同的形式。本章所介绍的这些常用的方差估计方法包括:复合抽样、对称重复抽样(BRR)、“折叠式”重复抽样(JRR)、自主抽样法和泰勒级数法。

# 第 1 节｜复合抽样：一种通用的方法

这种方法的根本原理是通过选取一系列复合子样本而不是单个样本来辅助方差计算。它要求每个子样本都用同一个选择方案独立地被选中。然后用同样的方法对每一个子样本进行估计，这样我们就可以通过这些独立的子样本估计值的变异情况，对基于任何子样本的总体估计的抽样方差进行估计。这与第 2 章中重复系统抽样的原理一样。

对于参数 $U$ 的 $t$ 个复合样本估计值 $u_1$，$u_2$，…，$u_t$ 的均值($\bar{u}$)，其抽样方差可以通过下面这个简单的方差方程估计出来(Kalton，1983：51)：

$$v(\bar{u}) = \sum (u_i - \bar{u})^2 / t(t-1) \qquad [4.1]$$

这个估计可以用于所有抽样设计的独立复合样本的所有样本统计值。

在应用这个估计方法时，Deming(1960)建议对描述性统计值使用 10 个复合样本，其他人(Sudman，1976)则建议最少用 4 个。当复合样本的数量在 3—13 之间时，标准误差大致的估计值可以计算为等于复合样本估计值的幅度除以复合样本的数量(Kish，1965：620)。但是，因为在统计推论中，$t$ 个复合样本的方差估计基于 $t-1$ 个自由度，因此在分析

性研究中我们需要更多的复合样本,数目大约在20—30之间(Kalton, 1983: 52)。

为了充分理解复合抽样方案的策略,我们来看一个简单的例子。假设我们想估计200个新生婴儿中男孩的比例。我们用Cochran(1977: 19)书中的随机数字模拟这个调查,假设奇数代表男孩。从该书表中的前10列,我们选择了10个复合样本,每个样本的样本规模 $n = 20$。每个复合样本中男孩的数量如下:

| 每个复合样本中男孩的数量 | 9 | 8 | 13 | 12 | 14 | 8 | 10 | 7 | 10 | 8 | 总计 = 99 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 男孩的比例 | 0.45 | 0.40 | 0.65 | 0.60 | 0.70 | 0.40 | 0.50 | 0.35 | 0.50 | 0.40 | 比例 = 0.495 |

男孩的总体百分比为49.5%,其标准误差为 $\sqrt{49.5 \times 50.5/200} = 3.54\%$。用方程4.1从这10个复合样本估计值中估计出来的标准误差为3.58%。要得到一个约为3.50%的大致估计很简单,只要用这些估计值的幅度(70% − 35%)除以复合样本的数量10即可。复合抽样方法的主要优势在于很容易对标准误差进行估计。

在实际中,选择独立复合样本的根本原则在某种程度上并不严格。例如,复合样本的选择使用的是无放回抽样而不是有放回抽样。对于不等概率设计,基本权重的计算以及无应答和事后分层的调整通常只对全样本执行一次,而不会对每一个复合子样本都分别进行这些计算和调整。在集群抽样中,研究者通常用与集群最初被选中时相同的次序系统地把不同的集群分配给 $t$ 个复合样本,这样可以充分利用分层效应。在使用方程4.1时,从全样本中得到的样本均值通常

用做复合样本均值的平均数。这些偏离根本原则的做法会影响方差估计,但在大型调查中,这种偏差应该是可以忽略的(Wolter, 1985: 83—85)。

作为美国国立精神卫生研究院1984年ECA调查的一部分,在康涅狄格州纽黑文县举行的社区精神卫生调查(E. S. Lee、Forthofer、Holzer & Taube, 1986)是复合抽样的一个典型的例子。这个调查的抽样框是按地理位置排列的居住电网清单。这个调查把两个居住单位作为一个集群,随机选取起始点,以61户为间距进行系统抽样。样本中的这一连串集群随后按顺序分配给12个子样本。这些子样本用于在长时间的筛选和访问中辅助数据的调度和中期分析。其中10个子样本用于社区调查,其他2个用于其他分析。这10个复合样本在此用于说明方差估计的程序。

这些子样本并没有严格遵循独立复合抽样的基本原则,因为除了第一个随机的起始点,其他起始点都是系统地选取的。但是,在这个例子中,系统地把集群分配给子样本可以产生一个近似的分层,从而得到更加稳定的方差估计,也因此比对这些相对少量的复合样本使用随机起始点更加可取。

所以,我们把这些子样本当做复合样本,并用方程4.1的复合抽样方差估计法进行估计。

通过Kish随机选择表(Kish, 1949),每一个家庭里面有一个成年人被随机选中,因此每个家庭里面成年人的数量就是每个观察值的样本权重。随后对这个权重进行无应答和事后分层调整。样本权重的创建只对全样本而不会对每一个子样本分别进行。这即为在分析中使用的权重。

**表 4.1 复合样本中标准误差的估计:纽黑文县 ECA 调查,1984( $n = 3058$ )**

| | | | 回归置信区间[a] | | | |
|---|---|---|---|---|---|---|
| 复合样本 | 流行率[b] | 数 比[c] | 截 距 | 性 别 | 种 族 | 年 龄 |
| 全样本 | 17.17 | 0.990 | 0.2237 | −0.0081 | 0.0185 | −0.0020 |
| 1 | 12.81 | 0.826 | 0.2114 | 0.0228 | 0.0155 | −0.0020 |
| 2 | 17.37 | 0.844 | 0.2581 | 0.0220 | 0.0113 | −0.0027 |
| 3 | 17.87 | 1.057 | 0.2426 | −0.0005 | 0.0393 | 0.0015 |
| 4 | 17.64 | 0.638 | 0.1894 | 0.0600 | 0.2842 | −0.0029 |
| 5 | 16.65 | 0.728 | 0.1499 | 0.0448 | −0.0242 | −0.0012 |
| 6 | 18.17 | 1.027 | 0.2078 | −0.0024 | −0.0030 | −0.0005 |
| 7 | 14.69 | 1.598 | 0.3528 | −0.0487 | −0.0860 | −0.0028 |
| 8 | 17.93 | 1.300 | 0.3736 | −0.0333 | −0.0629 | −0.0032 |
| 9 | 17.86 | 0.923 | 0.2328 | −0.0038 | 0.0751 | −0.0015 |
| 10 | 18.91 | 1.111 | 0.3008 | −0.0007 | 0.0660 | −0.0043 |
| 范围 | 6.10 | 0.960 | 0.2237 | 0.1087 | 0.3702 | 0.0038 |
| 标准误差 | | | | | | |
| 复合样本 | 0.59 | 0.090 | 0.0234 | 0.0104 | 0.0324 | 0.0004 |
| SRS | 0.68 | 0.097 | 0.0228 | 0.0141 | 0.0263 | 0.0004 |

注:a. 因变量(1 为状态存在,0 为状态不存在)对性别(男 = 1,女 = 0),肤色(黑人 = 1,非黑人 = 0)和年龄做回归。这个分析只是作为演示而已。

b. 过去 6 个月有任何一种精神疾病的百分比。

c. 6 个月流行率的性别差异。

资料来源:取自"Complex Survey Date Analysis: Esimation of Standard Errors Using Pseudo-Strata," E. S. Lee, Forthofer, Holzer, and Taube, *Journal of Economic and Social Measurement*. 1986 年版权所有。

表 4.1 展示了对全样本以及各复合子样本计算出来的三种统计值。(精神障碍)流行率的方差估计值,用百分数 ($p$) 的形式表示,可以利用方程 4.1 从复合样本估计值($p_i$)中计算出来:

$$v(p) = \frac{\sum (p_i - 17.17)^2}{10(10-1)} = 0.3474$$

其标准误差为$\sqrt{0.3474}=0.59$。由于应答率不同，总体流行率17.17%与10个复合样本估计值的均值稍微不同。我们注意到，复合样本估计值幅度的1/10[①]，即0.61与从方程4.1中得到的标准误差很接近。类似地，还可以对比数比(odds ratio)和回归系数估计标准误差。这些标准误差的估计值与从假定简单随机抽样(用教科书中的合适的方程)而计算出来的值几乎相同。这意味着对这个调查中的这些统计值来说设计效应非常小。

虽然复合抽样设计为我们提供了一种易于计算的方差估计方法，但为了获得合格的统计推论的精确度，还必须要有足够数量的复合样本才行。然而，如果复合样本的数量很多而每个样本的样本量却相对较小，就会严重限制在每个复合样本里面使用分层。更重要的是，在复杂抽样设计中使用复合抽样是很不切实际的。因此，复合抽样设计很少在大型的分析调查中使用。相反，复合抽样的原理常在数据分析阶段被用于估计方差。这种尝试促使了方差估计的拟复合抽样方法的出现。下面要讲的就是基于这种拟复合抽样的原理而产生的两种方法。

① 幅度除以样本数10。——译者注

# 第2节 | 对称重复抽样

对称重复抽样(BRR)是复合抽样原理在配对选择设计中的应用,其中每一个阶层内都有两个初级抽样单位(PSUs)被选取。配对选择设计代表了分层的最大化使用并允许对方差进行计算。在这种情况下,两个抽样单位之间的方差等于他们之间差异的平方的一半。为了应用复合抽样原理,我们首先把样本分成随机的组别以形成拟复合样本。如果是分层的设计,需要所有的阶层都能在每个拟复合样本中出现。在一个分层的配对选择设计中,我们只能建立两个拟复合样本:第一个从每个阶层的两个抽样单位中选出其中一个;每个阶层中剩下的另一个抽样单位则组成第二个拟复合样本(补充样本)。每个拟复合样本几乎都包含了总样本的一半。

把 $t=2$ 代入方程4.1,我们可以对这两个复合样本估计值 $u'$ 和 $u''$ 的均值的抽样方差进行估计,通过:

$$v(\bar{u}) = [(u' - \bar{u})^2 + (u'' - \bar{u})^2]/2 \qquad [4.2]$$

如方程4.2所示,复合样本估计值的均值经常用从全样本中获得的总体估计值代替。但是这个估计值太不稳定,以至于没有任何的实际价值,因为它只基于两个拟复合样本。

BRR则通过重复建立半样本的复合样本和从不同的阶层中选取不同的抽样单位解决了这个问题。这些拟复合的半样本因此包含了某些共同的单位，这样就导致了复合样本间的相关，从而使估计变得复杂了。其中一个解决方法——这种方法对线性统计值得出的方差估计是无偏的——是通过使用正交矩阵平衡拟复合样本的构造来实现的（Plackett & Burman，1946）。充分的平衡需要矩阵的大小等于4的倍数，而且复合样本的数量要大于或等于阶层的数量。这样某个样本统计值的抽样方差就可以估计为方程4.2中得到的方差估计值的均值除以$t$个复合样本数：

$$v(\bar{u}) = \sum[(u'_i - \bar{u})^2 + (u''_i - \bar{u})^2]/2t = \sum(u'_i - u''_i)^2/4t \qquad [4.3]$$

通过去掉补充的半样本的复合样本，可以把方程简化为：

$$v'(\bar{u}) = \sum(u'_i - \bar{u})^2/t \qquad [4.4]$$

这个估计方法最初是由McCarthy（1966）提出的。McCarthy阐释了这种平衡的方法以产生线性估计的无偏方差估计值。对非线性估计来说，方差估计是有偏差的，但很多研究证明这种偏差很小。在阶层数目较大的情况下，通过使用更少量的部分对称复合样本（K. H. Lee，1972；Wolter，1985：125—130），还可以进一步简化这种计算方法。

就像在复合抽样中一样，BRR假定初级抽样单位（PSU）是在阶层内通过有放回抽样而获得的，虽然实际操作中通常用的是无放回抽样。理论上，当应用到无放回样本中时，会导致方差的过高估计，但这种过高估计其实可以忽略不计，

因为在抽样比例很小时,无放回抽样方案选取同一个单位超过一次的机会很低。配对选择设计中抽样比例(假定在 BRR 方法中)通常很小,因为从每个阶层内只选取了两个 PSU。

当用于多阶段选择设计中时,BRR 通常只应用到 PSU 而无视 PSU 内部的次级抽样。这是必定的,因为当第一阶段的抽样比例很小时,抽样方差可以充分地从 PSU 之间的变异情况大概估算出来。这被称为最终集群逼近法。正如 Kalton(1983:第 5 章)所描述的,对简单两阶层选择设计的方差的无偏估计包括从这两个阶层中各得到的一个组成部分,但从第二个阶层中得到那个组成部分要乘以第一个阶层的抽样比例。因此,随着第一个阶层抽样比例的减少,第二个阶层的作用可以忽略不计。这种只基于 PSU 的快捷的计算方法在为复杂数据做公开准备以及对这种数据进行分析时显得尤为方便,因为除了第一阶段的抽样信息,并不需要这些复杂设计的其他详细的特征信息。

如果 BRR 要用于配对选择设计之外的其他方案,就有必要对数据结构进行调整以适应这种方法。在许多多阶段调查中,分层的数目被最大化了,最终只从每个阶层中选取一个 PSU。在这种情况下,PSU 之间可以进行配对以形成合并层从而使用 BRR。这种方法通常会导致对方差的某种过高估计,因为某些阶层间的变异性现在被包含在阶层内的计算结果之中了。对线性统计值来说,如果这种合并是明智而审慎地进行的,那么这个问题就不是很严重;但是,对非线性统计值,通常不建议使用这种合并方式(Wolter,1985:48)。稍后介绍的泰勒级数法可以用于非线性统计值的估计。还有一种对每个阶层内包含的三个 PSU 建立正交平衡的方

法，但它并没有被广泛使用(Gurney & Jewett, 1975)。

现在我们把BRR这种方法应用到1984年的美国GSS。正如前一节所述，它用的是多阶段选择设计。第一阶段的抽样从84个阶层——郡级地区或郡级地区群——中各选取一个PSU。前面的16个阶层为大都市地区，被定为具有自我代表性(或者自动包含在样本之中)。为了使用BRR，这84个阶层被合并成42对拟阶层。因为数据中不具自我代表性的PSU的编号服从阶层的地理排序，因此配对是以PSU的编码为基础按顺序进行的。之后，再把这16个具自我代表性的阶层合并成8个拟阶层，而剩下的68个不具自我代表性的阶层则被合并成34个阶层。但是，对这些具自我代表性的阶层进行配对的方式错误地包括了它们之间的异质性。为了排除这些，而仅仅只包括每个具自我代表性的阶层内部的异质性，这些在每个具自我代表性的拟阶层内部的观察值被结合起来再随机地分成了两个拟初级抽样单位。

为了平衡从42个拟阶层中产生的半样本的复合样本，我们使用了一个44阶正交矩阵(见表4.2)。这个矩阵由数字0和1构成，为了和这42个阶层匹配，前面两列被舍弃了(即44行对应复合样本，42列对应拟阶层)。0表示包括阶层中的第一个PSU，1表示包含第二个PSU。行是复合样本，列是阶层。比如，第一个复合样本包含了42个拟阶层中每个阶层的第二个PSU(因为第一行所有的值都等于1)。利用正交矩阵的行，44个复合样本和44个补充复合样本就产生了。

为了对从全样本中产生的统计值进行方差估计，我们首先需要对这44个复合样本和补充复合样本分别计算出我们关注的统计值。在计算这些复合样本的估计值时，我们使用

**表 4.2 44 阶正交矩阵**

| 行 | 列(44) |
|---|---|
| 1 | 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 |
| 2 | 1 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 |
| 3 | 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 |
| 4 | 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 |
| 5 | 1 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 |
| 6 | 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 |
| 7 | 1 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 |
| 8 | 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 |
| 9 | 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 |
| 10 | 1 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 |
| 11 | 1 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 |
| 12 | 1 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 |
| 13 | 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 |
| 14 | 1 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 |
| 15 | 1 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 |
| 16 | 1 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 |
| 17 | 1 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 |
| 18 | 1 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 |
| 19 | 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 |
| 20 | 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 0 0 1 1 |
| 21 | 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 |
| 22 | 1 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 |
| 23 | 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 |
| 24 | 1 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 |
| 25 | 1 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 |
| 26 | 1 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 |
| 27 | 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 0 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 |
| 28 | 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 |
| 29 | 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 |
| 30 | 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 |
| 31 | 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 |
| 32 | 1 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 |
| 33 | 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 0 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 1 |
| 34 | 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 |
| 35 | 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 |
| 36 | 1 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 0 |
| 37 | 1 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 0 |
| 38 | 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1 |
| 39 | 1 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 |
| 40 | 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 |
| 41 | 1 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 |
| 42 | 1 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 0 |
| 43 | 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 |
| 44 | 1 1 0 0 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 |

资料来源:经出版社同意,取自 Wolter(1985:第 32 题)。

**表4.3 BRR复合样本中同意一个成年人攻击另一个成年人的比例估计值：美国综合社会调查，1984($n = 1473$)**

| | 估计值(%) | | | 估计值(%) | |
|---|---|---|---|---|---|
| 重复次数 | 复合样本 | 补充复合样本 | 重复次数 | 复合样本 | 补充复合样本 |
| 1 | 60.9 | 59.2 | 23 | 61.4 | 58.6 |
| 2 | 60.1 | 59.9 | 24 | 57.7 | 62.4 |
| 3 | 62.4 | 57.9 | 25 | 60.4 | 59.6 |
| 4 | 58.5 | 61.7 | 26 | 61.7 | 58.2 |
| 5 | 59.0 | 61.0 | 27 | 59.3 | 60.6 |
| 6 | 59.8 | 60.2 | 28 | 62.4 | 57.6 |
| 7 | 58.8 | 61.5 | 29 | 61.0 | 58.9 |
| 8 | 59.0 | 61.0 | 30 | 61.2 | 58.7 |
| 9 | 61.3 | 58.8 | 31 | 60.9 | 59.1 |
| 10 | 59.2 | 60.8 | 32 | 61.6 | 58.5 |
| 11 | 61.7 | 58.3 | 33 | 61.8 | 58.2 |
| 12 | 60.2 | 59.8 | 34 | 60.6 | 59.4 |
| 13 | 62.1 | 58.7 | 35 | 58.6 | 61.5 |
| 14 | 59.7 | 60.4 | 36 | 59.4 | 60.7 |
| 15 | 58.1 | 62.0 | 37 | 59.8 | 60.3 |
| 16 | 56.0 | 64.2 | 38 | 62.0 | 58.1 |
| 17 | 59.8 | 60.3 | 39 | 58.1 | 61.9 |
| 18 | 58.6 | 61.3 | 40 | 59.6 | 60.5 |
| 19 | 58.6 | 61.1 | 41 | 58.8 | 61.2 |
| 20 | 60.8 | 59.3 | 42 | 59.3 | 60.8 |
| 21 | 63.4 | 56.5 | 43 | 58.7 | 61.4 |
| 22 | 58.3 | 61.7 | 44 | 60.5 | 59.5 |

总体估计 = 60.0

| 方程估计 | 方差 | 标准误 | 设计效应 |
|---|---|---|---|
| 基于方程4.3 | 0.000231 | 0.0152 | 1.42 |
| 基于方程4.4 | 0.000227 | 0.0151 | 1.40 |

了调整后的抽样权重。表4.3显示了对这44个复合样本和他们的补充复合样本分别估计出的同意"攻击"的成年人的比例。总体比例为60.0%。由方程4.3估计出来的总体比例的抽样方差为0.000231。把这个结果与SRS设计中该比

例的抽样方差[$pq/(n-1)=0.000163$，忽略 FPC]进行比较，可知设计效应等于 1.42(=0.000231/0.000163)。这个设计效应意味着从 GSS 中得到的比例估计值的方差比从具相同样本规模的 SRS 中得到的方差大 42%。由方程 4.4 得到的方差估计值也类似。

综上所述，BRR 使用了拟复合抽样的方法来估计抽样方差，它主要是为配对选择方案而设计的。它也可以用于某些通过配对阶层而只从每个阶层中选择一个 PSU 的复杂的调查，但是这种配对方法必须是明智而审慎的，而且要考虑到实际的样本选择程序。在大多数可用的 BRR 的应用软件包中，对于复合样本中被选中的 PSU，其观察值的抽样权重被增加至两倍，以弥补没有被选中的那一半 PSU。此外，Fay 认为，在创建复合权重时还存在一个 BRR 的变量，用 $2-k$ 还是 $k$ 乘以初始权重，主要取决于以正交矩阵($0\leqslant k<1$)为基础的 PSU 是被选中了还是没有被选中。这将在下一章进一步说明。

# 第3节 |"折叠式"重复抽样

"折叠式"方法最初作为一种估计偏差的非参数程序，是由 Quenouille(1949)提出的，随后 Tukey(1958)介绍了相同的这种程序怎样用于估计方差。Durbin(1959)在他的开创性比率估计研究中首先使用了这种方法。之后，Frankel(1971)采用与 BRR 相同的方式，把它应用于复杂调查中的方差计算，而且还把这种方法命名为"折叠式"重复抽样法。跟 BRR 一样，JRR 通常用于阶层内的 PSU。

我们可以通过对一个简单随机样本的样本均值进行抽样方差估计来阐释"折叠式"方法的基本原理。假设 $n=5$，$y$ 的样本取值为 3，5，2，1 和 4。那么样本均值 $\bar{y}=3$，它的抽样方差(忽略 FPC)等于：

$$v(\bar{y}) = \frac{\sum (y_i - \bar{y})^2}{n(n-1)} = 0.5 \qquad [4.5]$$

均值的"折叠式"方差的产生如下：

(1) 删除第一个样本值，然后计算出拟样本均值，即 $\bar{y}_{(1)} = (5+2+1+4)/4 = 12/4 = 3$。接着，换成删除第二个样本值，得到第二个拟均值 $\bar{y}_{(2)} = 10/4$；以此类推，$\bar{y}_{(3)} = 13/4$，$\bar{y}_{(4)} = 14/4$，$\bar{y}_{(5)} = 11/4$。

(2) 对这 5 个拟均值计算它们的均值 $\bar{\bar{y}} = \sum \bar{y}_{(i)}/n = (60/4)/5 = 3$，这个均值与样本均值相等。

(3) 然后方差就可以从这 5 个拟均值(每个都包含 4 个观察值)的变异性中被估计出来：

$$v(\bar{\bar{y}}) = \frac{(n-1)\sum(\bar{y}_{(i)} - \bar{\bar{y}})^2}{n} = 0.5 \qquad [4.6]$$

这个结果与方程 4.5 的结果相同。

这些基于复合抽样的程序有一个明显的优势:它们可以同时应用于那些无法用方程表达出来的估计值,如样本中位数,以及那些以公式为基础的估计值。中位数的抽样方差没有可用的公式,但是“折叠式”程序却可以为其提供一个估计值。用上面这个例子,样本中位数是 3,而 5 个对应的拟中位数分别为 3，2.5，3.5，3.5 和 2.5(这些拟中位数的均值为 3)。用方程 4.6,中位数的方差估计为 0.8。

以同样的方式,“折叠式”程序还可以应用于复合抽样。我们可以每次去掉其中一个复合样本并计算出相应的拟值来估计“折叠式”方差,虽然在这种情况下它并没有任何计算优势。但是它也可以应用于任意由概率抽样构成的随机组别。比如,为使用“折叠式”程序,一个系统抽样样本可以被分成随机或系统的子样本。对其他抽样设计来说,可以按照 Wolter(1985：31—33)提出的实际规则构成随机组别。基本原理是,随机组别以下述方式产生:每个随机小组都具有与母样本相同的抽样设计。这需要实际抽样设计详细的信息，但这样的信息在公开的调查数据中通常是得不到的。因此，“折叠式”程序通常用于 PSU 而不是随机组别。

对于配对选择设计,复合样本是通过——从一个阶层中

去掉一个 PSU,然后再对剩下的那个 PSU 加权以保持这个阶层在总样本中的比例——这种方式构成的。补充复合样本的构成方式是一样的,只需要把这个阶层中去掉的和保留的 PSU 对换一下,对每个复合样本都估计一个拟值。对于加权后的样本,保留下来的 PSU 的抽样权重需要加大以弥补那些被去掉的 PSU 里面的观察值。放大后的权重等于保留下来的 PSU 中所有权重的和除以一个因子 $(1 - w_d/w_t)$,其中 $w_d$ 是被去掉的 PSU 中所有权重的和,而 $w_t$ 是该阶层内所有 PSU 的权重的和。这个因子代表了被删除的 PSU 占总阶层权重的比例的补集。①然后,在配对选择设计中,某个样本统计值的方差就可以由阶层 $h$ 中拟值 $u'_h$ 和补充拟值 $u''_h$ 通过下列方程计算出来:

$$v(\bar{u}) = \sum[(u'_h - \bar{u})^2 + (u''_h - \bar{u})^2]/2 = \sum(u'_h - u'')^2/4 \qquad [4.7]$$

这个估计值与方程 4.3 的形式一样,可以调整为每个阶层只包含一个复合样本(不与补充样本结合起来取均值),像 BRR 中方程 4.4 一样,得出:

$$v'(\bar{u}) = \sum(u'_h - \bar{u})^2 \qquad [4.8]$$

JRR 并不局限于配对选择设计,它可用于每个阶层含任意数量 PSU 的情况。如果我们用 $u_{hi}$ 表示第 $h$ 个阶层第 $i$ 个复合样本的 $U$ 的估计值,$n_h$ 表示第 $h$ 个阶层中被抽中的 PSU 的数量,$r_h$ 表示在第 $h$ 个阶层中形成复合样本的数量,那么

---

① 即 1 减被删除的 PSU 占总阶层权重的比例。——译者注

方差就可以通过下面的方程估计而得：

$$v(\bar{u}) = \sum_{h}^{L}\left(\frac{n_h - 1}{r_h}\right)\sum_{i}^{r_h}(u_{hi} - \bar{u})^2 \qquad [4.9]$$

如果阶层 $h$ 内每个 PSU 都依次被删掉以形成不同的复合样本，那么在每个阶层内 $r_h = n_h$，但并不需要在第 $h$ 个阶层构造 $n_h$ 个复合样本。当阶层的数量很大而 $n_h$ 大于或等于 2 时，计算方程可以简化为只使用每个阶层内的 1 个复合样本。但是，在分析性研究中，必须使用足够多的复合样本以保证有足够的自由度。

表 4.4 给出了把 JRR 应用到在 BRR 计算中使用过的 1984 年美国 GSS 合成配对设计中的结果。对 42 个“折叠式”复合样本和它们的补充样本，分别列出了“同意攻击其他成年人”的成年人比例的估计值。利用方程 4.7，我们得出方差估计值为 0.000238，设计效应为 1.46，这个结果与用 BRR 得出的结果相差不大。把补充样本去掉只包含 42 个复合样本时(方程 4.8)，方差估计值等于 0.000275，设计效应为 1.68。

**表 4.4 JRR 复合样本中同意一个成年人攻击另一个成年人的比例估计值：美国综合社会调查，1984**

| | 估计值(%) | | | 估计值(%) | |
|---|---|---|---|---|---|
| 重复次数 | 复合样本 | 补充复合样本 | 重复次数 | 复合样本 | 补充复合样本 |
| 1 | 60.2 | 59.8 | 22 | 60.3 | 60.0 |
| 2 | 60.2 | 59.8 | 23 | 60.0 | 60.0 |
| 3 | 60.0 | 60.0 | 24 | 60.4 | 59.6 |
| 4 | 60.3 | 59.8 | 25 | 60.1 | 59.8 |
| 5 | 60.0 | 60.1 | 26 | 59.8 | 60.3 |
| 6 | 59.9 | 60.1 | 27 | 59.9 | 60.1 |
| 7 | 60.0 | 60.0 | 28 | 60.1 | 60.0 |
| 8 | 60.0 | 60.0 | 29 | 59.5 | 60.3 |

续表

| | 估计值(%) | | | 估计值(%) | |
|---|---|---|---|---|---|
| 重复次数 | 复合样本 | 补充复合样本 | 重复次数 | 复合样本 | 补充复合样本 |
| 9 | 59.9 | 60.2 | 30 | 59.9 | 60.1 |
| 10 | 60.1 | 60.0 | 31 | 59.6 | 60.2 |
| 11 | 59.8 | 60.2 | 32 | 60.5 | 59.6 |
| 12 | 59.9 | 60.1 | 33 | 60.1 | 59.9 |
| 13 | 59.8 | 60.2 | 34 | 60.3 | 59.8 |
| 14 | 60.0 | 60.1 | 35 | 60.1 | 59.8 |
| 15 | 59.6 | 60.5 | 36 | 60.2 | 59.8 |
| 16 | 60.4 | 59.6 | 37 | 60.0 | 60.0 |
| 17 | 59.9 | 60.0 | 38 | 59.6 | 60.4 |
| 18 | 59.8 | 60.2 | 39 | 59.9 | 60.1 |
| 19 | 59.8 | 60.2 | 40 | 60.5 | 59.6 |
| 20 | 59.9 | 60.1 | 41 | 60.4 | 59.8 |
| 21 | 60.0 | 60.0 | 42 | 60.7 | 59.4 |

总体估计 = 60.0

| 方差估计 | 方差 | 标准误 | 设计效应 |
|---|---|---|---|
| 基于方程 4.7 | 0.000238 | 0.0152 | 1.46 |
| 基于方程 4.8 | 0.000275 | 0.0166 | 1.68 |

仔细观察表 4.4，我们可能会觉得 JRR 复合估计值的变异程度没有表 4.3 中 BRR 复合估计值的变异程度那么大。但我们要注意，JRR 代表一种使用了不同方法估计方差的不同的策略。注意 BRR 中方程 4.3 的分母包含了复合样本的数量 $t$，而 JRR 中方程 4.7 则与复合样本的数量无关。原因在于，在 JRR 中，复合估计值本身已经与复合样本的数量相关。由于那些复合样本的产生过程已经删除了其中一个单位，因此与只有少量的抽样单位可用于构造复合样本的情况相比，当可供利用的抽样单位数目很大时，得出的复合估计值会更加接近于总体估计值。因此，不需要在方程 4.7 和方程 4.8 中加入复合样本的数量。但是，当复合样本的数量比

PSU的总数量少的时候,就需要把复合样本的数量考虑进来,如方程4.9所示。

综上所述,JRR以拟复合抽样方法为基础,可以用于估计复杂抽样调查中的抽样方差。不需要对样本选择方案做任何限制,但产生复合样本时要非常小心,必须把最初的抽样设计考虑在内。正如前面提到的,这些详细的设计信息对二手分析研究者来说很难得到。比如,如果GSS数据文件可以提供更多的关于终端集群的信息,我们就可以构造更加方便而且更接近真实抽样设计的随机组群,而不需要把JRR用于折叠配对设计中。

# 第4节 | 自主抽样法

与BRR和JRR密切相关的另一种方法是由Efron(1979)推广的自主抽样法。其基本原理是通过重复地从观察到的数据中抽取初级抽样单位(PSU)来建立具相同规模且与设计中的结构相同的复合样本。把自主抽样法应用到GSS数据中的42个拟阶层中的84个PSU中,我们会抽到84个PSU(使用有放回抽样方法),每两个从同一个阶层中得来。在某些阶层中,相同的PSU可能会被选择两次。这种抽样方法常常要重复多次,最少200(标记为B)次(Efron & Tibshirani, 1993:第6.4节)。但是,如要对更少的变量进行估计,通常就需要多非常多的复合样本(Korn & Graubard, 1999: 33),对每个新建立的复合样本($u_i'$)都进行参数估计。然后,所有复合样本估计值的均值,其方差的自主抽样估计值表示为:

$$v(\bar{u}) = \frac{1}{B}\sum_{i=1}^{B}(u_i' - \bar{u})^2 \qquad [4.10]$$

我们需要把这个估计值乘以 $(n-1)/n$ 以纠正偏差。当$n$很小时,偏差可能非常大。在我们的例子中,每个阶层内有两个PSU,估计到的方差需要平分。另一个纠正偏差的方法是在阶层$h$内重新抽取($n_h-1$)个PSU,然后把这些抽中的

PSU 内的观测值的抽样权重乘以 $n_h/(n_h-1)$ (Efron, 1982: 62—63)。在我们的例子中,这样会产生如 BRR 中的半样本复合样本。基于至少 200 个复合样本的自主抽样估计与基于 44 个半样本复合样本的 BBR 估计差不多相同。由于自主抽样法需要大量的复制,因此这种方法并没有广泛应用于复杂调查分析的方差估计。

学者们已经提出了把自主抽样法应用到方差估计中的各种不同的方法(Kovar、Rao & Wu, 1988)。虽然基本的原则广为人知,但在选择自主抽样样本时却有许多不同的方法。比如,Chao 和 Lo(1985)建议把主样本中的每个观察值复制 $N/n$ 次以产生无放回简单随机抽样的自主抽样总体。对那些不等概率抽样方案来说,观察值的复制需要与抽样权重成比例;也就是说,要用 PPS 选择自主抽样样本。目前,还没有详尽的研究分析这些可供选择的方法以及可能出现的因为偏离独立同分布样本这个基本假设而带来的影响。

虽然自主抽样法对处理许多统计问题非常有用,但对于复杂调查中的方差估计来说,它不如 BRR 和 JRR 实用,因为它需要大量的复合样本。对不同的使用者,BRR 和 JRR 会产生相同的结果,但自主抽样法可能会因不同的使用者或者相同使用者的不同使用方式而出现不同的结果,因为复制程序就很可能每次都产生不同的结果。正如 Tukey(1986: 73)所述,与自主抽样法和其他模拟方法相比,“目前折叠法似乎是最实际可行的用于估计众多不确定性来源的方法”。目前自主抽样法并没有被安装在现有的复杂调查分析的软件包中,尽管它已被广泛用于统计计算的其他领域。

# 第5节 | 泰勒级数法(线性化)

泰勒级数扩展法已经应用于数学和统计的许多不同情况。它的一个早期的应用是用于获取难以计算的函数值的近似值,比如指数 $e^x$ 或者对数[log(x)]函数。它出现在计算机能完成该项特殊功能之前,以及在我们还无法获得相关的适用表时。$e^x$ 的泰勒级数扩展涉及对 $e^x$ 求关于 $x$ 的一阶导数或者更高阶的导数;对某些值——通常是 0——估计这些导数;并在这些导数的基础上建立一些列(数列/方程)项。$e^x$ 的扩展为:

$$1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\frac{x^4}{4!}+\cdots$$

下面这个是以 $a$ 扩展的一般方程的应用:

$$f(x)=f(a)+f'(a)(x-a)+\frac{f''(a)(x-a)^2}{2!}+\frac{f'''(a)(x-a)^3}{3!}+\cdots$$

在统计上,泰勒级数法用于获取某些非线性函数的近似值。通常,近似值为函数提供了合理的估计,有时候近似值甚至是一个线性函数。这种方差估计的方法在现有文献中有几种不同的名称,包括线性化方法、delta 方法(Kalton,

1983：44)和方差传递法(Kish，1965：583)。

在统计应用中，扩展估计用的是 $x$ 的均值或者期望值 $E(x)$。如果我们用 $E(x)$ 代替上面那个一般方程中的 $a$，会有：

$$f(x) = f[E(x)] + f'[E(x)][x - E(x)] + f''[E(x)][x - E(x)]^2/2! + \cdots$$

根据定义，$f(x)$ 的方差 $V[f(x)] = E[f^2(x)] - E^2[f(x)]$，使用泰勒级数扩展法，会有：

$$V[f(x)] = \{f'[E(x)]\}^2 V(x) + \cdots \qquad [4.11]$$

对不止一个随机变量的函数，方法也一样。对具两个变量的函数，泰勒级数扩展法会产生：

$$V[f(x_1, x_2)] \cong \left(\frac{\partial f}{\partial x_1}\right)\left(\frac{\partial f}{\partial x_2}\right) Cov(x_1, x_2) \qquad [4.12]$$

把方程 4.12 应用到两个变量 $x$ 和 $y$ 的比——即 $r = y/x$——之中，我们会得到比率估计的方差方程：

$$V(r) = \frac{V(y) + r^2 V(x) - 2rCov(x, y)}{x^2} + \cdots$$

$$= r^2\left(\frac{V(y)}{y^2} + \frac{V(x)}{x^2} - \frac{2Cov(x, y)}{xy}\right) + \cdots$$

把方程 4.12 扩展到 $c$ 个随机变量的情况，$\theta = f(x_1, x_2, \cdots, x_c)$ 的近似方差为：

$$V(\theta) \cong \sum\sum\left(\frac{\partial f}{\partial x_i}\right)\left(\frac{\partial f}{\partial x_j}\right) Cov(x_i, x_j) \qquad [4.13]$$

把方程 4.13 应用到 $n$ 个观察值样本中的 $c$ 个变量的加权后估计中，

表 4.5 用泰勒级数法估计的标准误差:同意成年人攻击另一个成年人的百分比,美国 1984 年综合社会调查($n=1473$)

| | 子样本 | 估计值(%) | 标准误差(%) | 设计效应 |
|---|---|---|---|---|
| 总体 | | 60.0 | 1.52 | 1.41 |
| 性别 | 男 | 63.5 | 2.29 | 1.58 |
| | 女 | 56.8 | 1.96 | 1.21 |
| 种族 | 白人 | 63.3 | 1.61 | 1.43 |
| | 非白人 | 39.1 | 3.93 | 1.30 |
| 教育 | 大学 | 68.7 | 2.80 | 1.06 |
| | 高中毕业 | 63.3 | 2.14 | 1.55 |
| | 其他 | 46.8 | 2.85 | 1.27 |

$$f(Y)=\hat{Y}_i=\sum w_i y_{ij},\ j=1,\ 2,\ \cdots,\ c$$

Woodruff(1971)指出:

$$V(\theta)\cong V\Big[\sum w_i \sum\Big(\frac{\partial f}{\partial y_j}\Big)y_{ij}\Big] \qquad [4.14]$$

这种非线性估计的另一种线性方差形式具有计算上的优越性,因为它避免了方程 4.13 中 $c\times c$ 协方差矩阵的计算。一个简单的加法的互换就实现了把多阶段估计问题转化成单变量问题的便利。这种通用的计算程序可用于不同的非线性估计,包括回归系数(Fuller, 1975; Tepping, 1968)。

在复杂调查中,这种近似值法被应用到阶层内的 PSU 总体。即,方差估计值是同一阶层内不同 PSU 的(方程 4.14 中)方差的加权联合值。这些方程很复杂,但需要的计算时间远比上面我们讨论过的复制方法所用的时间少得多。这种方法可用于任何一种用数学方式表达的统计量,比如均值或者回归系数,但不能用于那些非函数式的统计量,如中位数或者其他百分位数。

我们现在来看 GSS 例子中样本比例的方差估计。表 4.5 列出了按性别、种族和教育程度分析,泰勒级数法应用于同意攻击别人的成年人的百分比时的结果。该比例等于所有正面回复的权重之和与所有权重的比。其标准误差的计算根据方程 4.14,做了相应修改以包括 PSU 和阶层。总体比例的设计效应为 1.41,这与用另外两种方法估计出来的结果(见表 4.3[BRR]和表 4.4[JRR])大体相同。估计到的比例随性别、种族和教育程度的不同而不同。由于子样本很小,所以子样本中的标准误差比总体中的要大。此外,子样本中的设计效应也与总体中的不一样。

这一章中,我们介绍了几种复杂调查中统计量的方差估计方法(更深入的介绍请参考 Rust & Rao, 1996)。GSS 和其他调查的例子表明,在大部分复杂调查中,设计效应大于 1。更多的例子,可以参考 E. S. Lee、Forthofer 和Lorimor(1986)以及 Eltinge、Parsons 和 Jang(1997)。第 6 章的例子也将说明在复杂调查数据分析中上述其中一种方法的重要性。

# 第5章

# 调查数据分析的准备

前几章重点关注的是调查设计的复杂性和这些设计的方差估计方法。在应用抽样权重和估计设计效应的方法之前,我们必须了解调查设计和对统计量进行相关估计的数据要求以及需要使用的软件。相比从 SRS 中得到的数据,对复杂调查数据来说这些要求在某种程度上更加严格。

# 第1节 | 调查分析的数据要求

正如第3章所述,权重和设计效应是恰当的调查数据分析所需的基本因素。在对从二手来源中得到的调查数据做分析准备工作时,除了感兴趣的变量,我们还必须把权重、(数据文件中的)抽样单位和阶层的识别变量也包括进来。由于在不同的调查数据源中这些与设计相关的条目其命名各不相同,因此很有必要仔细阅读相关文件或者向数据来源机构或个人咨询,以了解调查设计和数据准备程序。

在主要的调查数据来源中,权重通常是公开可用的。正如前面指出的,权重反映了选择概率以及无应答和事后分层调整。这些权重通常表示为扩展权重,扩展权重总和等于总体规模。在某些特定的分析中,把它们转化成相对权重后可能更加方便,相对权重总和等于样本规模。某些调查数据可能包含几个权重变量以辅助数据的恰当使用,这几个权重可能对应于不同的目标总体或者特定数据条目的子样本。通过仔细阅读相关文件,对不同的数据用途选择适当的权重是很有必要的。对某些调查来说,权重这个变量并没有明确标明,因此需要研究抽样设计以产生这些权重。如第4章所述,在GSS中,权重是从家庭中成年人的数量得出的。此外,也很有必要进行事后分层调整以使样本构成与总体构成一

致(如表 3.1)。如果在与数据提供方联系后仍然无法得到权重,使用者也不能假定数据具有自加权性质。如果没有权重就用数据,使用者必须在报告结果时明确说明这点。很难想象在没有权重的情况下对调查数据进行分析,即便是在基于模型的分析之中,这种情况我们也会意识到可能存在不等选择概率和子样本的不同应答率(虽然与基于设计的分析使用方法不同)。

设计效应的计算通常需要第一阶段选择程序方面的信息,即阶层和 PSU 的识别;在某些特定的嵌入式设计中,还需要次级抽样单位和相关的层方面的信息。如果从每个层内选择一个 PSU(如在 GSS 中),层的标识与 PSU 的标识是一样的。如果没有用到分层或者数据并没有提供层标识,那么我们可以假定其为无限制集群抽样,然后进行分析。如果没有任何阶层与 PSU 信息,那么就很有必要分析一下,对于给定的抽样设计,把数据当成 SRS 是否合理。

当使用了分层且层标识可供利用时,我们需要确保每个阶层内至少有两个 PSU,否则就不可能估计方差。如果每个阶层内只有一个集群被选中,就需要对层进行配对以产生拟阶层。对层进行配对需要充分理解抽样设计。第 3 章列举了 GSS 中一个特定的阶层配对策略的例子。当相关文档中没有任何有用的信息时,也可使用随机匹配的方法。基于美国国民健康调查,Stanek 和 Lemeshow(1977)研究了匹配的效应,发现加权后均值的方差估计和联合比估计对不同的阶层匹配方法没有什么反应,但这个结果可能并不是对所有的调查都适用。

# 第2节 | 预备性分析的重要性

调查数据分析以前期探索研究为起点，以检验数据是否适合用于有意义的分析。对二手数据来源的前期研究中，一个重要的考虑因素是要检验不同的子群体中是否存在足够的观察值以支持后面的分析。基于未加权的列表结果，分析者可以确定样本规模是否足够大，以及变量的不同类别是否需要合并。未加权的列表结果同时还能显示具缺失值或者极端值的观察值的数量，可反映是否存在测量误差或者录入错误。

虽然受访单位(受访者)无应答调整是由数据收集机构在创建抽样权重时进行处理的，但分析者必须自己处理缺失值，即选项无应答。如果缺失值较少，我们可以在分析中忽略那些具缺失值的受访者。但是，基于设计的调查分析通常不会完全排除具缺失值的观察值，这种分析通过把具缺失值的观察值的权重设为0而仍然使用完整的数据库。这对于估计抽样设计中固有的方差是非常有必要的。无论是把观察值完全排除还是把它的权重设为0，点估计的结果都一样，但方差估计的结果却不相同。把观察值完全排除会低估方差。

如果选项无应答情况比较严重，那么忽略缺失值的做法

就会把太多的权重设为 0,这样原来的加权方案就会遭到破坏从而产生偏差,也就不可能再准确反映目标总体。解决这个问题的方法是加大没有缺失值的观察值的权重,以弥补那些被忽略掉的观察值。在实施这种调整时,我们通常假定具缺失值的对象内部并不存在系统的模式,但这个假定也可能无效。比如,如果所有具缺失值的对象都是男性或者都属于某个特定的年龄群体,那么简单地加大那些剩下的观测值的权重就不合适。另一个可行的权重调整办法是利用某些合理的方法填补这些缺失值,但这个方法并不一定比加权调整的方法好。

缺失值的填补不是常规的统计工作。填补缺失值的方法有很多种(Kalton & Kasprszky, 1986; Little & Rubin, 2002)。在选择某种特定的填补方法之前,非常有必要先了解导致缺失值的机制。在某些情况下,我们可以使用简单的方法。比如,对类别变量来说,可以多加一个类别表示缺失值。对连续变量,另一个简单的办法是均值填补,但是这种方法会扭曲变量的分布。为了保持原分布,我们可以使用热卡填补、回归填补,或者多重填补,下一章我们将对填补方法进行简单(并不深入的)说明。如果使用了填补法,方差估计就可能需要某些调整(Korn & Graubard,1999:第 5.5 节),但这个问题并不在本书的讨论范围之内。

在进行任何实质性分析之前,很有必要检验一下是否每个 PSU 都有足够的观察值。由于无应答和缺失值的排除,某些 PSU 很可能只包含非常少的观察值,甚至根本没有观察值。一个没有观察值或者只有很少观察值的 PSU 可以与同一阶层内相邻的 PSU 合并起来。由于合并 PSU,因此只

有一个 PSU 的阶层会与相邻的阶层合并起来。但是，合并太多的 PSU 和阶层会破坏原来的抽样设计。随之而来的数据分析也可能出现问题值，因为已经不可能再判断(PSU/阶层)合并后的样本代表的究竟是哪个总体。每个 PSU 所需要的观察值的数量取决于所要进行的分析类型。分析性研究比描述性统计量的估计所需要的观察值数量更大。普遍的法则是这个数量应该足够大，以对某个估计值计算 PSU 内部方差。

为了说明这点，我们来看看 GSS 数据。按阶层和 PSU 进行的未加权列表结果显示，PSU 内部观察值的数量在 8 至 49 之间，大部分大于 13，这意味着 PSU 很可能已经足够大，可用于估计均值和百分比的方差。在分析性研究中，我们可能需要分析按教育程度和性别分类后同意攻击(别人)的成年人的百分比。对于这个分析，我们需要检查在特定的分性别教育类别内，是否存在一些没有任何观察值的 PSU。如果对某些分性别教育类别存在许多没有观察值的 PSU，这就会使方差—协方差矩阵的估计(这个估计基于阶层内 PSU 总体的变异情况)出现问题。按 PSU 得出的分性别(两类)教育(三类)列表结果显示，84 个 PSU 中有 42 个至少有一个分性别教育单元没有观察值。即使把教育程度合并成两类，我们也需要合并几乎一半的 PSU。因此，我们不应就关于攻击这个问题同时分析性别和教育这两个变量与攻击问题的关系，但可以在不需要合并较多 PSU 和阶层的情况下，单独分析性别或者教育与攻击问题的关系。

我们不能通过在分析范围内选取观察值的方法进行复杂调查数据的子群体分析。虽然个案选择不会改变基本权

重,但它可能会破坏基本的抽样设计。比如,选择一个小的种族类别可能会去掉一部分 PSU,剩下的 PSU 中观察值的数量也会显著减少。结果难以从选到的子集中评估基本设计中固有的设计效应。即使基本设计并没有完全被破坏,但从复杂调查样本中选择观察值也可能导致方差的不准确估计,正如上面处理缺失值的方法中所描述的一样。准确的方差估计需要在分析中保持完整的数据并对在分析范围外的观察值赋予 0 权重。软件包中可用的子总体分析方法,其子群体分析并不对分析范围内的观测值进行选择。我们将在第 6 章对子总体分析做进一步介绍。

前期分析的第一步是检查主要变量的基本分布。列表可以指出是否需要修正变量的操作定义以及是否需要对特定变量合并其取值类别。基于汇总统计,我们可以知道样本中某些变量的有意思的模式和分布情况。在对变量进行逐一分析之后,我们接下来可以检验变量间是否存在关系,以排除那些并不相关的变量或者与因变量无关的变量。我们可以用标准的、以 SRS 为基础的图形统计方法进行前期检验。但是,由于调查数据中权重的重要作用,任何忽略了权重前期分析的检验都不可能达到预期目标。一种把权重考虑在内的前期分析方法是选择具可操作规模的子样本,其选择概率与权重大小成比例,以利用标识统计图形方法分析这个子样本。第 6 章第一部分将对这个方法进行介绍。

# 第3节 | 方差估计方法的选择

把设计特征结合到分析中去需要对方差估计方法进行选择。如第4章所述，在实际中有三种方差估计方法(BRR、JRR和泰勒级数近似法)。不同的研究者(Bean, 1975; Frankel, 1971; Kish & Frankel, 1974; Lemeshow & Levy, 1979)对这三种方法进行了经验性评估，Krewski 和 Rao (1981)还对这些方法做了一些理论上的比较。这些评估研究显示，这三种方法中没有哪一种能始终比其他两种更好或者更坏，在多数情况下这些方法的选择可能取决于软件的适用性或者对软件的熟悉情况。少数情况下，这些方法的选择可能取决于需要进行估计的统计量的类型或者使用的抽样设计，就像在配对选择设计中一样。

基于公式的泰勒级数近似法可能是应用最广泛的复杂调查的方差估计方法，因为它在大部分可用的软件中都能找到。由于现实原因，它可能比基于复制的方法(BRR 和 JRR)更好，但正如第4章所述，它不适用于中位数或其他百分位数以及非参数统计量。基于复制的方法更普遍，且可应用于这些统计量，但它们需要创建和处理复合样本。基于复制的方法的另一个优点是它提供了一个简单的方法以更加合理地融入无应答和事后分层调整。通过单独对每个复合样本

进行加权调整,它可以在方差估计中融入调整的效应。

对小型调查和小范围估计来说,JRR 可能比 BRR 更稳定,因为在 JRR 中每个复合样本都包含了大部分全样本,而 BRR 复合样本只包含一半的样本。但对于百分位数估计,BRR 比 JRR 更加可靠(Kovar et al., 1988; Rao、Wu & Yue, 1992)。Fay 建议的 BRR 的一种变体(Judkins,1990)被用于稳定方差估计。在这个变体中,复合权重被加大了,被乘以$(2-k)$或$k$而不是 2 或者 0;而且方差估计也被调整了,即把方程 4.4 的右边乘以$1/(1-k)^2$($k$的取值在 0 和 1 之间)。Korn 和 Graubard(1999: 35—36)发现,当$k=0.3$时,Fay 的 BRR 会产生与标准 BRR 差不多相同的结果,但当把无应答和事后分层调整结合到复合权重中时,会得到稍微小一点的方差。Fay 的方法可以视为是 JRR 和 BRR 的折中。Judkins(1990)发现,对于百分位数和其他统计量的估计,当$k=0.3$时,Fay 的方法在偏差和稳定性方面比标准 BRR 或者 JRR 都更好。

BRR 的设计目的是用于配对选择设计。当我们从每个阶层中选择一个 PSU 时,必须对这些 PSU 进行匹配以创建拟阶层从而使用 BRR 方法。当从每个阶层中选出来的 PSU 的个数大于 2 时,就难以建立配对设计,使用 JRR 或者泰勒级数近似法会更好。从程序上说,泰勒级数近似法是最简单的,而基于复制的方法需要更多的步骤以创建复合权重。

# 第4节 可用的计算资源

过去的30年,有许多不同的程序被开发出来用以分析复杂调查数据。早期的程序在政府机构、调查研究组织和大学中因不同的目的而开发出来。其中某些程序进一步开发成针对普通用户的程序包。最初的程序包是为主机计算应用而开发出来的。随着个人电脑计算能力的增强,更多的努力投向了PC版本的开发,一些新的软件包也随之出现。虽然其中一些已经不再更新,但还有三种程序包与软件标准的现有状况保持一致,而且便于使用。它们是SUDAAN、Stata和WesVar。

SUDAAN软件包已经流行了20多年,可从北卡罗来纳的三角研究所获取。它有两个版本:单机版和SAS(统计分析系统)版。后者结合SAS特别便于使用。与SAS一样,SUDAAN的许可证需要每年更新一次。

这种软件的方差估计的默认方法是泰勒级数近似法,也可选择使用BRR或JRR。它几乎可以处理所有类型的抽样设计,包括多层嵌套和事后分层设计。它具有可用于复杂调查数据分析的最全面的分析特征,但相比另外两个软件包,它的维护需要更多的费用。它支持一系列不同的统计程序,包括CROSSTAB、DESCRIPT、RATIO、REGRESS、

LOGISTIC、LOGLINK(对数线性回归)、MULTILOG(多类别和次序 logistic 回归)和 SURVIVAL(Cox 比例风险模型)。过去的这些年,设计者加进了一些新的程序,去掉了一些旧的程序,如用于加权最小二乘建模的 CATAN。SAS 使用者可能会觉得这些程序很容易执行,但对初学者来说,如果没有有经验用户的帮助,可能难以明确该选用哪种设计,也难以解释输出结果。SAS 网站提供了咨询和技术支持。

Stata 是包含了调查分析模板的通用的统计程序包,可从德克萨斯州大学城 Stata 公司获取。它的调查分析部分支持一系列不同的分析性程序,包括 svymean、svytotal、svyprop(比例)、svyratio(比率估计)、svytab(二维表)、svyregress(回归)、svylogit(logistic 回归)、svymlogit(多类别 logistic 回归)、svypois(poisson 回归)、svyprobit(probit 模型)以及其他。它利用 PSU 用泰勒级数近似法进行方差估计(最终集群逼近法)。虽然它并不支持复杂的设计如多层嵌套设计,但可用于分析大部分实用的调查设计。其大部分调查分析程序与它的通用(非调查)统计程序类似,这意味着,它的统计分析中的许多普遍特征可轻易与调查分析部分结合起来。它的输出结果也相对比较容易理解,因此新手可能会觉得 Stata 比 SUDAAN 更容易学。

WesVar 程序是由美国马里兰州罗克维尔市的 Westat 公司(Westat, Inc., Rockville, Maryland)研发和推广的。它利用复制的方法计算描述性统计量、线性回归建模和对数线性模型。有 5 种可用的复制方法,包括 JK1(用于无分层设计的 delete-1 jackknife)、JK2(用于每层两个设计的 jackknife)、JKn(用于分层设计的 delete-1 jackknife)、BRR 和 Fay(使用

Fay 的方法的 BRR)。它由 Westat 公司研发(Flyer & Mohadjer, 1988),此软件包的旧版本可免费获取。现在这个软件包在市场上可以买到,而且还有一个学生版本。虽然可从其他系统中导入数据,但这个软件包最初的设计模式是单机版软件包。除了需要全样本的抽样权重,它还要求输入数据文件中的每条记录都包含复合权重。对简单设计来说,它可以在运行任何程序之前先创建复合权重。关于这个软件的文档信息是很充足的,但如果没有有经验用户的帮助,初学者可能会觉得关于创建复合权重的描述有点难以理解。

Cohen(1997)就编程难度、效率、精确度和程序设计能力等方面,评估了这三种软件包的早期版本(SUDAAN 版本 7, Stata 版本 5, WesVarPC 版本2.02)。评估结果显示,WesVar 程序一贯要求用最少的编程命令得出需要的调查估计值,但它要求准备附加数据以创建推导方差估计所必需的复合权重。Stata 需要更多的命令来得出同样的结果,但它不会在执行命令时产生过分的压力。就计算效率来说, SUDAAN 在产生所需的估计值时始终占优势。

很难说哪一种软件包比其他软件包更值得推荐。在选择软件包时,我们应该在分析需要的背景下,考虑要用到的方差估计方法、维持这个软件的费用以及之前所讲到的这些软件包各自的优点和缺陷。更重要的一点也许是分析者对软件包的熟悉程度。比如对 SAS 用户来说,选择 SUDAAN 是很自然的。Stata 用户也很可能会选择使用 Stata 的调查分析部分。第 6 章对 Stata(版本 8)和 SUDAAN(版本 8.0.1)的使用做了详细说明,这些说明会为软件包的选择提供额外的建议。

随着更多的统计软件包融入复杂调查数据分析程序,计算资源的可用性一直在不断完善。SPSS 13.0 现在给出了一个附加模块用于调查数据分析——SPSS 复杂样本模块(SPSS Complex Samples)。它包含四个程序:CSDESCRIPTIVES、CSTABULATE(列联表分析)、CSGLM(回归,ANOVA,ANCOVA)和 CSLOGISTIC。SAS 系统也在最新版本 SAS9.1 中提供了调查数据的分析能力。SURVEYFREQ 程序能产生具相关检验的单向和多维列联表分析。SURVEYLOGISTIC 程序执行logistic 回归,也适用于其他联结函数。目前我们可以利用许多现成可用的统计软件包来进行调查数据分析,而不需要任何特设的软件。

# 第5节 | 创建复合权重

BRR方法需要复合权重。这些权重可能包含在数据库中或者在运行分析之前创建。通常需要创建 $H$(4的倍数,大于阶层的数量)组复合权重。对于SUDAAN来说,复合权重输入在数据之中,WesVar可以用适当的设置创建复合权重。比如,对于具六层且每层含两个PSU的调查来说,如表5.1所示,需要建立八组复合权重。这个例子中BRR的SUDAAN命令列在表格的左边[①]。输入的数据包括层,PSU,病床的数量,艾滋病人的数量,抽样权重和八组复合权重(w1到w8)。注意复合权重要么等于样本权重的2倍(如果这个单元被选中),要么等于0(如果这个单元没有被选中)。这些0或2倍的权重根据6行8×8的正交矩阵进行排列,与表4.2类似。艾滋病人总体的比率估计(再次参考注释[2])等于目标区域内艾滋病(床位)与床位总数(2501)的比。PROC RATIO和DESIGN = brr指定说明所要的统计量和估计方差的方法,deff要求得出设计效应。因为使用的是BRR设计,所以不需要NEST这个命令(指定说明阶层和PSU)。REPWGT指定复合权重变量。NUMER和DENOM指定说明比率估

① 原文中为"右边",有误。——译者注

**表 5.1　BRR 复合权重的创建和 Jackknife 程序**

(A) BRR 的 SUDAAN 命令

```
data brr;
input stratum psu beds aids wt w1-w8;
aids2=aids*2501;
datalines;
1 1  72 20 2  0  0  0  4  0  4  4  4
1 2  87 49 6 12 12 12  0 12  0  0  0
2 1  99 38 2  4  0  0  0  4  0  4  4
2 2  48 23 2  0  4  4  4  0  4  0  0
3 1  99 38 2  4  4  0  0  0  4  0  4
3 2 131 78 4  0  0  8  8  8  0  8  0
4 1  42  7 2  0  4  4  0  0  0  4  4
4 2  38 28 2  4  0  0  4  4  4  0  0
5 1  42 26 2  4  0  4  4  0  0  0  4
5 2  34  9 2  0  4  0  0  4  4  4  0
6 1  39 18 4  0  8  0  8  8  0  0  8
6 2  76 20 2  4  0  4  0  0  4  4  0
;
proc ratio design=brr deff;
weight wt;
repwgt wi-w8;
numer aids2;
denom beds;
run;
```

BRR 的 SUDAAN 输出结果

```
The RATIO Procedure
Variance Estimation Method: BRR
by: Variable, One.
```

| 变量 | | One 1 |
|---|---|---|
| AIDS2/BEDS | Sample Size | 12 |
| | Weighted Size | 32.00 |
| | Weighted X-Sum | 2302.00 |
| | Weighted Y-Sum | 2741096.00 |
| | Ratio Est. | 1190.75 |
| | SE Ratio | 154.15 |
| | DEFF Ratio #4 | 2.10 |

(B) 刀切法的 SUDAAN 命令

```
Data jackknife;
input stratum psu beds aids wt;
aids2=aids*2501;
datalines;
1 1  72 20 2
1 2  87 49 6
2 1  99 38 2
2 2  48 23 2
3 1  99 38 2
3 2 131 78 4
4 1  42  7 2
4 2  38 28 2
5 1  42 26 2
5 2  34  9 2
6 1  39 18 4
6 2  76 20 2
;
proc ratio design=jackknife deff;
nest stratum;
weight wt;
numer aids2;
denom beds;
run;
```

刀切法的 SUDAAN 的输出结果

```
The RATIO Procedure
Variance Estimation Method: Delete-1 Jackknife
by: Variable, One.
```

| 变量 | | One 1 |
|---|---|---|
| AIDS2/BEDS | Sample Size | 12 |
| | Weighted Size | 32.00 |
| | Weighted X-Sum | 2302.00 |
| | Weighted Y-Sum | 2741096.00 |
| | Ratio Est. | 1190.75 |
| | SE Ratio | 141.34 |
| | DEFF Ratio #4 | 1.76 |

资料来源:Levy and Lemeshow(1999:384)。

计的分子和分母。由 SUDAAN 而得的 BRR 估计值见表 5.1 左下部分。估计值等于 1191,其标准误差为 154.2(设计效应为 2.10)。同样的数据,WesVar 基于同一正交矩阵不同的行来创建复合权重,其得出的点估计是一样的,而标准误差为 155.0。使用不同的复合权重(抽样权重的 1 或 0 倍)的 BRR 得出的标准误差为 149.5。Fay 的 BRR 变体使用的复合权重根据某个观察值是否在被选中的单元内而等于抽样权重的 $2-k$ 或者 $k$ 倍($0\leqslant k<1$)。当 $k=0.3$ 时,标准误差估计为 137.3。

Jackknife 程序并不需要复合权重。这个程序通过在每个复合样本中删除一个 PSU 来创建复合样本。JRR 的

SUDAAN命令和结果见表5.1的右部分。由Jackknife程序估计到的标准误差为141.3，比BRR估计的结果小。由泰勒级数近似法（假定为有放回抽样）得到的标准误差为137.6，比Jackknife估计值稍微小点，而与Fay的BRR估计结果类似。如第4章所述，BRR和JRR假定有放回抽样。如果我们假定无放回抽样（使用有限总体校正法），这个例子中的标准误差估计值为97.3。

美国国家卫生统计中心（NCHS）的第三次国民健康与营养调查（NHANES III）[4]包含了BRR的复合权重。这些复合权重是为Fay的方法而创建的，$k=0.3$，且融入抽样不同阶段的无应答和事后分层调整。如Korn和Graubard（1999）提出的，使用Fay的把无应答和事后分层调整融入在内的创建复合权重的办法更可取。但许多调查数据库通常没有这种权重，也缺少适当的创建这种复合权重的信息。

# 第6节 寻找合适的调查数据分析的模型

据说许多统计分析是在没有确立清晰的目标的基础上开展的。在分析数据之前,仔细考虑研究问题并建立一个明确的分析计划是很重要的。如前面章节所述,在调查分析中对数据进行前期分析和探索是非常重要的。相比基于设计的分析,这个任务在基于模型的分析中要困难得多。

其中可能会涉及提出问题或者开展合适的背景研究以获取选择某个合适的模型所需要的信息。调查分析者通常没有参与数据收集,要理解数据收集的设计通常很困难。对最初的设计进行提问可能并不充分,但却很有必要弄清楚设计方案如何在实地调查中执行。通常,相关的设计信息既无法在相关文档内找到也无法从数据库中得到。再者,给定可能的样本规模,某些调查设定的目标可能也太过不切实际。所谓的通用调查不可能包括后来使用者关心的所有问题。要建立一个包含所有相关变量的、合适的模型是一个挑战。

此外还应该检查已有知识,尤其当类似的数据曾在以前被分析过时。最好不要从头开始拟合模型,相反可以检查一下新的数据是否与先前的(研究)结果一致。遗憾的是,在社会科学和健康科学研究中很难找到使用复杂调查数据的基

于模型的分析。许多处理基于模型分析的文章关注的是在理想化的条件下分析调查数据的最优程序。比如，大部分公共调查数据库只包含阶层和 PSU、多层或者分层线性模型(Bryk & Raudenbush，1992；Goldstein & Silver，1989；Korn & Graubard，2003)界定额外目标参数的机会有限。复杂调查数据分析的混合线性模型的使用需要进一步的研究，我们希望它能促使调查设计者把设计和分析更紧密地结合起来。

# 第6章

# 调查数据分析的操作

这一章会对调查数据分析做各种说明,重点介绍融入权重和数据结构对分析的影响。我们首先介绍一种对大型复杂调查进行前期分析的策略。我们利用 NHANES III 阶段 II 的数据对各种不同的分析加以说明,包括描述性分析、线性回归分析、列联表分析和 logistic 回归分析。对每种分析,我们都将讨论调查数据要求的一些理论上或实际上的考虑因素。分析中所使用的变量的选择是基于阐释方法的需要,并不是为了说明实质性的发现。最后,我们还将就基于模型的视角进行讨论,因为它与本章列举的分析性例子相关。

# 第1节 | 预备性分析的策略

抽样权重可能会在复杂调查数据的前期分析中造成破坏，但在数据探索的过程中忽略权重并不是令人满意的解决办法。另一方面，用于调查数据分析的程序并不适用于基本的数据探索。而且图解法并不是专门与复杂调查一起设计出来的。在这一部分，我们会介绍一种把权重考虑在内的前期分析策略。

在发明计算机之前，权重的处理在数据分析中有各种不同的方法。当IBM分类机器用于数据列表时，常用的做法是复制数据卡片以匹配权重值从而获取合理的估计值。为了加快大型调查的制表速度，某些调查采用了PPS方法(Murthy & Sethi, 1965)。意识到分析复杂调查数据的困难性，Hinkins、Oh和Scheuren(1994)提倡使用逆抽样设计演算法，这种方法可从现有的复杂调查数据中产生一个简单随机子样本，用户因此可以直接把传统的统计方法应用到这个子样本中。这些方法对调查数据分析已经不再有吸引力了，因为现在已经有现成可用的调查分析程序。但是，由于前期分析没有必要使用完整的数据库，因此用PPS方法选取子样本的原理仍然是为前期分析建立样本的一个好方法。

PPS子样本可用常规的描述性和图表方法进行分析，因

为权重已经反映在子样本的选择之中。比如,散点图是前期数据探索的其中一种最重要的图表法。一种把权重结合到散点图中的方法是使用代表权重大小的气泡。Korn 和 Graubard(1998)检验了各种不同的绘制双变量数据散点图的方法,说明了使用 PPS 子样本的优点。实际上,他们发现抽样散点图比气泡散点图更可取。

为了前期分析,我们从 NHANES III 阶段 II(1991—1994)(参见注释[4])的成年人数据(共 9920 个成年人)中产生了一个样本规模为 1000 的 PPS 样本。

**表 6.1 某些选定特征的子样本和全样本估计结果:美国 NHANES III 阶段 II 调查成年人人口**

| | 平均年龄 | 维他命的使用 | 西班牙人口 | SBP[a] | 样本 BMI[b] 和 SBP 的相关性 |
|---|---|---|---|---|---|
| 总样本($n$=9920)[c] | | | | | |
| 未加权 | 46.9 岁 | 38.4% | 26.1% | 125.9 mmHg | 0.153 |
| 加权后 | 43.6 | 42.9 | 5.4 | 122.3 | 0.243 |
| PPS 子样本($n$=1000) | | | | | |
| 未加权 | 42.9 | 43.0 | 5.9 | 122.2 | 0.235 |

注:a. 收缩压。
b. 体重指数。
c. 大于或等于 17 岁的成人。

我们首先按阶层和 PSU 对总样本进行分类,然后在累积相对权重的取值范围内用 9.92 的跳跃间距系统地选取 PPS 子样本。按阶层和 PSU 进行的分类整理从根本上保留了原有抽样设计的完整性。

表 6.1 显示了可用传统统计软件包进行分析的 PPS 子样本的实用性。我们选择了几个最受权重影响的变量。由于我们对老年人和少数种族进行了过取样,平均年龄和西班

牙人口比例的加权后估计值与未加权估计值不一样。此外，权重还使维他命使用情况和收缩压的估计结果不一样，因为它们受到过抽样类别的严重影响。虽然没有经过加权，但子样本中的估计值却与加权后的总样本中的估计值非常接近，显示了前期分析中 PPS 子样本的有用性。在体重指数与收缩压的关联度这方面也存在类似的结果。

PPS 子样本在并不正式考虑权重而对数据进行探索时非常有用，对还处于入门阶段的学生尤其有用。它特别适合通过如散点图、对比箱形图和中位数标示图等图形方法探索数据。其真正的优势在于重新抽样过的数据近似地代表了总体，可以在忽略权重的情况下进行分析。从这种重新抽样过的数据中得到的点估计几乎与完整数据库中的加权后估计值相等。从这种重新抽样过的数据中发现的有趣的趋势很可能会被更完整的(通过 Stata 或者 SUDAAN 进行的)分析所证实，虽然标准误差可能会有所不同。

# 第 2 节 | 描述性分析

我们使用 NHANES III 阶段 II 中的成年人样本(17 岁或以上)进行描述性分析。这个样本包含 9920 个观察值,这些观察值处于 23 个拟阶层中,每个阶层含 2 个拟初级抽样单位。拟阶层和初级抽样单位的标识包括在我们的工作数据文件之中。数据文件中扩展权重被转化为相对权重。为了检查在不同的 PSU 内其观察值的分布是否存在问题,我们制作了未加权列表。结果显示,每个 PSU 中包含的观察值的数量在 82 至 286 之间。这些 PSU 的样本规模对进一步分析已经足够。

我们决定分析体重指数、年龄、种族、贫困指数、收缩压、维他命使用情况和抽烟状况。体重指数等于体重(公斤)除以高度(米)的平方。年龄用年测量,教育用受教育年限测量,贫困指数等于家庭收入与贫困程度的比,收缩压用 mmHg测量。此外,我们还选择了以下几个二分类变量:Black(1=黑人;0=非黑人),Hispanic(1=西班牙裔;0=非西班牙裔),使用维他命(1=是;0=否),抽烟状况(1=曾经抽过;0=从未抽过)。

对分析中选中的变量我们填补了其缺失值以说明调查数据分析的步骤。在以往的研究中已有各种不同的填补办

法用以弥补缺失的调查数据(Brick & Kalton, 1996; Heitjan, 1997; Horton & Lipsitz, 2001; Kalton & Kasprszky, 1986; Little & Rubin, 2002; Nielsen, 2003; Zhang, 2003)。也有几种软件包可供利用(如 SAS/STAT 中的 proc mi 和 proc mianalyze、SOLAS、MICE、S-Plus Missing Data Library)。有很多方法可把这些软件包应用到某个特定的数据库中。合适的应用方法和(处理)过程的选择最终取决于缺失值的数量、产生缺失值的机制(是否可以忽略)以及缺失值的模式(单调的还是一般的)。复杂统计方法的使用很吸引人,但却可能弊大于利。更好的方法是根据具体的实例(Kalton & Kasprszky, 1986; Korn & Graubard, 1999:第 4.7 节和第 9 章),而不是技术性的操作指引来选择合适的方法。关于这方面详细的讨论不在本章讨论之列。以下简单的介绍仅供解说之用。

在我们的数据中,年龄和种族没有缺失值。我们先填补那些缺失值最少的变量值。维他命使用情况和抽烟状况的缺失值少于 10 个,教育和高度的缺失值大约为 1%。我们使用热卡填补[5](就近填补)的方法对这 4 个变量填补缺失值——按性别在五年一组的年龄类别内以与样本权重成比例的概率随机选取捐赠观察值。当某个观察值内还有更多的变量具缺失值时,我们也使用相同的观察值对缺失值进行填补。我们使用回归填补法来填补高度(3.7%缺失;基于权重、年龄、性别和种族)、权重(2.8%缺失;基于高度、年龄、性别和种族)、收缩压(2.5%缺失,基于高度、权重、年龄、性别和种族)、贫困指数(10%缺失,基于家庭规模、教育和种族)等的缺失值。填补的贫困指数值约 0.5%为负值,这些值被

设成0.001(数据中最小的贫困指数值)。附带说明一下,我们可以把其他人体测量的方法放到回归填补法之中,但我们的演示仅基于这个分析所选择的变量。最后,体重指数值(5.5%缺失)以更新后的权重和高度信息为基础重新计算而得。

为了说明抽样权重和设计效应会导致不同的结果,我们采用三种不同的方法进行分析:(1)未加权的,忽略数据结构;(2)加权后的,忽略数据结构;(3)调查分析,结合权重和抽样特征。第一种方法假定简单随机抽样,第二种考虑了权重但忽略了设计效应,第三种则为给定的抽样设计提供合适的分析。

首先,我们在填补和未填补缺失值这两种情况下,检验了加权后均值和比例以及它们的标准误差。填补的结果对点估计并没什么重大影响,用第三种分析法估计到的标准误差稍微变小了一点。没有填补缺失值时,加权后的贫困指数均值为3.198(标准误差 *s. e.* = 0.114),填补缺失值后结果为3.168(*s. e.* = 0.108)。对于体重指数,没有填补缺失值时,加权后的均值为25.948(标准误差 *s. e.* = 0.122),填补缺失值后结果为25.940(*s. e.* = 0.118)。其他变量,点估计值及其标准误差在这两种情况下得到的结果到小数点第三位相同,因为其缺失值很少。

估计到的描述性统计值(使用了填补值之后)列于表6.2。计算使用的是Stata。表格顶部未加权的统计值是由非调查分析命令 summarize(用于点估计)和 ci(用于标准误差)得出的。表格顶部加权后的估计值是通过相同的非调查分析命令加上[w=wgt]得到的。第三种分析(把权重和设计特

征融入进来)的结果列于表格底部。它使用 svyset[pweight =wgt], strata(stra)和 psu(psu)设定复杂调查特征,以及 svymean 对具体的变量估计均值和百分比。

表 6.2 显示的统计值为连续变量的均值估计值、二分类变量的百分比和标准误差估计。有几个变量其加权后和未加权的均值/百分比之间存在细微差别,但对某些变量来说,这种差别非常大。加权后黑人的比例比未加权时的比例小 60%以上,加权后西班牙裔的比例比未加权时的小将近 80%,反映了对这两个少数种族的过抽样。加权后的平均年龄约比未加权的平均年龄小 3.5 岁,原因在于对年老者进行了过抽样。另一方面,对贫困指数和受教育年限来说,加权后的均值比未加权的均值大很多,意味着被过抽样的少数种族集中于收入和教育分布中的底端。加权后的维他命使用情况的估计值也在某种程度上比未加权时大一些。这可能反映了少数种族较少使用维他命。

表格底部显示了(同时反映了权重和设计特征的)调查估计值。虽然估计到的均值和百分比与表格顶部加权后的结果完全相同,但对所有的变量来说,其标准误差都显著增大了。这种差别反映在表格中的设计效应中(表格底部标准误差与表格顶部加权后统计值的标准误差之比的平方)。贫困指数、教育和年龄的巨大的设计效应部分反映了这些特征的居住同质性。这两个社会经济特征变量(贫困指数和教育)和年龄的设计效应比黑人和西班牙裔这两个变量的设计效应要大。而 1976—1980 年的 NHANES II 中的结果却相反(数据用于本书第一版中),这意味着相比种族身份,现在在社会经济身份这方面居住区域更具同质性了。

**表 6.2 NHANES III 阶段 II 中 17 岁及以上成年人的回归分析中所用变量的描述性统计(用 Stata 进行的分析)**

(A) 加权和未加权的描述性统计，忽略设计因素

| Variable | Unweighted Analysis Mean | Std. Err. | Weighted Analysis Mean | Std. Err. | Min | Max |
|---|---|---|---|---|---|---|
| bmi | 26.4465 | .05392 | 25.9402 | .05428 | 10.98 | 73.16 |
| age | 46.9005 | .20557 | 43.5572 | .17865 | 17 | 90 |
| black | .2982 | .00459 | .1124 | .00317 | 0 | 1 |
| hispanic | .2614 | .00441 | .0543 | .00228 | 0 | 1 |
| pir | 2.3698 | .01878 | 3.1680 | .02086 | 0 | 11.89 |
| educat | 10.8590 | .03876 | 12.3068 | .03162 | 0 | 17 |
| sbp | 125.8530 | .20883 | 122.2634 | .18397 | 81 | 244 |
| vituse | .3844 | .00488 | .4295 | .00497 | 0 | 1 |
| smoker | .4624 | .00501 | .5114 | .00502 | 0 | 1 |

(B) 利用权重和设计因素的调查分析

```
. svyset [pweight=wgt], strata(stra) psu(psu)
. svymean bmi age black hispanic pir educat sbp vituse smoker

Survey mean estimation
pweight:  wgt                            Number of obs(*) =      9920
Strata:   stra                           Number of strata =        23
PSU:      psu                            Number of PSUs   =        46
                                         Population size  =   9920.06
```

| Mean | Estimate | Std. Err. | [95% Conf. | Interval] | Deff |
|---|---|---|---|---|---|
| bmi | 25.9402 | .11772 | 25.6946 | 26.2013 | 4.9903 |
| age | 43.5572 | .57353 | 42.3708 | 44.7436 | 10.3067 |
| black | .1124 | .00973 | .0923 | .1326 | 9.4165 |
| hispanic | .0543 | .00708 | .0397 | .0690 | 9.6814 |
| pir | 3.1680 | .10779 | 2.9622 | 3.4328 | 25.6331 |
| educat | 12.3068 | .12312 | 12.0565 | 12.5671 | 15.0083 |
| sbp | 122.2634 | .38543 | 121.4010 | 122.980 | 4.1995 |
| vituse | .4295 | .01215 | .4043 | .4546 | 5.9847 |
| smoker | .5114 | .01155 | .4874 | .5352 | 5.2829 |

注：* 有些变量有缺失值。

表格底部还显示了均值和百分比的 95%置信区间。置信限度所用的 $t$ 值并不是我们熟悉的可从 $n = 9920$ 样本中期望得到的 1.96(相对权重的和)。其原因在于，在多阶段集群抽样设计中，自由度是基于 PSU 和阶层的数量，而不是像 SRS 中那样基于样本规模。需要特别强调的是，复杂调查中自由度等于选中的 PSU 的数量减去所使用的阶层的数量。在我们的例子中，自由度为 23(= 46 − 23)，$t_{23,\,0.975} = 2.0687$，这个 $t$ 值用于表 6.2 中的所有置信区间。在某些情况下，自由度可能会与从上述一般方法中得到的稍微不同(Korn & Graubard，1999：第 5.2 节)。

在表6.3中，我们演示了进行子群体分析的例子。正如前一章所提到的，任何使用复杂调查数据进行的子群体分析都应该使用完整的数据库，而不能在分析范围内任意选取数据。在Stata中有两种合适的子群体分析的方法：使用by或者subpop。表6.3显示了对黑人这个变量进行子群体分析的例子。在表格顶部，我们用by分别对非黑人和黑人估计了BMI均值。黑人的BMI均值比非黑人的大。虽然在非黑人中BMI的设计效应(5.5)与总体的设计效应(表6.2中的5.0)相似，但在黑人中却只有1.1。

Stata还可用于检验参数的线性组合。两个子群体之均值是否相等可以用lincom命令进行检验(如[bmi]1－[bmi]0，检验关于当black = 1和当black = 0[①]时各自的总体BMI均值之间的差别的假设)，基于$t$检验这两个均值之间的差异在统计上是显著的。

设计效应是1.46，表示这个检验的$t$值降低了约20%以抵消抽样设计特征。

另一种方法，subpop可用于估计黑人的BMI均值，如表格底部所示。这种方法使用了完整的数据库，把不在分析范围内的观察值的权重设为0。均值、标准误差和设计效应与表格顶部用by命令得到的黑人的结果相同。

下面，我们通过明确(分析)范围(if black = 1)把黑人选择出来，从而估计BMI均值。因为在某些PSU中并没有黑人，所以这个方法并不奏效。按阶层和PSU的列表显示，在第13个和第15个阶层内只有一个PSU。当这两个阶层与

① 原文为nonblack = 0，有误。——译者注

**表 6.3 NHANES III 阶段 II 中 17 岁及以上黑人和非黑人平均体重指数的比较 ($n = 9920$)：用 Stata 进行的分析**

```
(A)  . svyset [pweight=wgt], strata(stra) psu(psu)
     . svymean bmi, by (black)

     Survey mean estimation
     pweight:  wgt                                      Number of obs    =      9920
     Strata:   stra                                     Number of strata =        23
     PSU:      psu                                      Number of PSUs   =        46
                                                        Population size  =   9920.06
     --------------------------------------------------------------------------------
     Mean     Subpop. |  Estimate     Std. Err.   [95% Conf. Interval]        Deff
     -----------------+--------------------------------------------------------------
     bmi     black==0 |   25.7738      .12925     25.5064     26.0412         5.512
             black==1 |   27.2536      .17823     26.8849     27.6223         1.071
     --------------------------------------------------------------------------------

(B)  . lincom [bmi]1-[bmi]0, deff

      ( 1) - [bmi]0 + [bmi]1 = 0.0
     --------------------------------------------------------------------------------
            Mean |  Estimate    Std. Err.      t     P>|t|   [95% Conf. Interval]   Deff
     ------------+-------------------------------------------------------------------
             (1) |   1.4799      .21867      6.77    0.000    1.0275     1.9322    1.462
     --------------------------------------------------------------------------------

(C)  . svymean bmi, subpop(black)

     Survey mean estimation
     pweight:  wgt                                      Number of obs    =      9920
     Strata:   stra                                     Number of strata =        23
     PSU:      psu                                      Number of PSUs   =        46
     Subpop.:  black==1                                 Population size  =   9920.06
     --------------------------------------------------------------------------------
         Mean |  Estimate     Std. Err.   [95% Conf. Interval]        Deff
     ---------+----------------------------------------------------------------------
          bmi |   27.2536      .17823     26.8849     27.6223         1.071
     --------------------------------------------------------------------------------

(D)  . svymean bmi if black==1
     stratum with only one PSU detected

(E)  . replace stra=14 if stra==13
     (479 real changes made)
     . replace stra=16 if stra==15
     (485 real changes made)

     . svymean bmi if black==1

     Survey mean estimation
     pweight:  wgt                                      Number of obs    =      2958
     Strata:   stra                                     Number of strata =        21
     PSU:      psu                                      Number of PSUs   =        42
                                                        Population size  =  1115.244
     --------------------------------------------------------------------------------
         Mean |  Estimate     Std. Err.   [95% Conf. Interval]        Deff
     ---------+----------------------------------------------------------------------
          bmi |   27.2536      .17645     26.8867     27.6206         2.782
```

邻近的阶层合并起来时，Stata 分析产生了结果。虽然点估计与前面的相同，但标准误差和设计效应却不同。普遍的法则是，调查数据的子群体分析应该避免筛选出子数据库，这与 SRS 数据的分析不同。

除了 svymean 命令外，Stata 还支持以下几种描述性分析：svytotal（用于总体估计）、svyratio（用于比率估计）、svyprop（用于百分比估计）。在 SUDAAN 中，这些描述性统计量可以用 DESCRIPT 程序进行估计，次范围分析则可通过 SUBPOPN 命令实现。

# 第 3 节 | 线性回归分析

回归分析和 ANOVA 都检验一个连续因变量和一系列自变量之间的线性相关。为了检验假设,假定因变量服从正态分布。下面这个方程表示了这些方法所考虑的相关方式。对于 $i = 1, 2, \cdots, n$,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i \qquad [6.1]$$

这是一个线性模型,因为因变量 $Y_i$ 由 $\beta_j$ 的线性组合加上 $\varepsilon_i$ 表示。$\beta_j$ 是方程中自变量 $X_j$ 的系数,$\varepsilon_i$ 是模型中的随机误差项——假定服从均值为 0 方差为常数的正态分布且相互独立。

在回归分析中,自变量要么是连续的,要么是离散的,而 $\beta_j$ 就是这些自变量对应的系数。在 ANOVA 中,自变量 $X_j$ 是指示变量(在效果编码下,自变量的每个类别都分别有一个编码为 1 或者 0 的指示变量),以显示哪种(因素的)效应被加入到模型中,而 $\beta_j$ 就是这些效应。

当数据来自 SRS 时,普通最小二乘估计用于获得回归系数的估计值或线性模型中的效应。但在处理从复杂抽样中得到的数据时,就需要对这种方法进行一些修改。现在我们用的数据包含个体观察值加上抽样权重和设计描述变量。

正如第3章所讨论的，复杂抽样中个体被选中的概率常常不同。此外，由于抽样设计方面的原因，在复杂调查中随机误差项通常不再相互独立。正因为这些与SRS偏离的情况，模型参数的OLS估计及其方差估计是有偏差的。因此，置信区间和假设检验就可能存在误导性。

很多学者已经讨论过这些问题(Binder, 1983; Fuller, 1975; Holt、Smith & Winter, 1980; Konijn, 1962; Nathan & Holt, 1980; Pfeffermann & Nathan, 1981; Shah、Holt & Folsom, 1977)，他们并没有统一使用某种单一的分析方法，但都认为OLS这种估计方法可能(在某些情况下)不合适。我们在此并不详细回顾所有前人的相关研究，而只重点关注一种可用于最广泛的情况之中的、有现成软件可供利用且已被广泛应用的方法。这种模型参数估计方法是设计加权最小二乘法(DWLS), SUDAAN、Stata和其他复杂调查数据分析软件都支持这种方法的使用。

DWLS方法中的权重是第3章所讨论的抽样权重。DWLS与不等方差的加权最小二乘法稍微不同，它是从假定的协方差结构中得到权重的(见Lohr, 1999:第12章)。考虑到由抽样设计和其他对权重的调整所带来的复杂性，第4章所讨论的其中一种方法可用于在模型参数估计值的方差—协方差矩阵估计中。由于这些模型使用的是PSU总体而不是个体值作为方差计算的基础，因此这个设计的自由度等于PSU的数量减去阶层的数量而不是样本规模。与误差平方和相关的自由度则等于PSU的数量减去阶层的数量再减去模型中公式项的数量。

表6.4显示了用前面提到的三种分析方法对BMI进行

**表 6.4 NHANES III 阶段 II 中成年人体重指数的多元回归模型 ($n = 9920$)：用 Stata 进行的分析**

(A) 不加权和加权分析，忽略设计因素

| Source | Unweighted analysis SS | df | MS | Weighted analysis SS | df | MS |
|---|---|---|---|---|---|---|
| Model | 33934.57 | 9 | 3770.48 | 37811.46 | 9 | 4201.27 |
| Residual | 252106.39 | 9910 | 25.44 | 236212.35 | 9910 | 23.84 |
| Total | 286040.68 | 9919 | 28.84 | 274023.81 | 9919 | 27.63 |

| | Unweighted analysis | Weighted analysis |
|---|---|---|
| F( 9, 9910) | 148.21 | 176.26 |
| Prob > F | 0.0000 | 0.0000 |
| R-squared | 0.1186 | 0.1380 |
| Adj R-squared | 0.1178 | 0.1372 |
| Root MSE | 5.0438 | 4.8822 |

| bmi | Coef. | Std. Err. | t | P>\|t\| | Coef. | Std. Err. | t | P>\|t\| |
|---|---|---|---|---|---|---|---|---|
| age | .38422 | .01462 | 26.27 | 0.000 | .39778 | .01528 | 26.03 | 0.000 |
| agesq | -.00391 | .00014 | -27.61 | 0.000 | -.00421 | .00016 | -27.06 | 0.000 |
| black | 1.15938 | .13178 | 8.80 | 0.000 | .96291 | .16108 | 5.98 | 0.000 |
| hispanic | .70375 | .14604 | 4.82 | 0.000 | .64825 | .22761 | 2.85 | 0.004 |
| pir | -.14829 | .03271 | -4.53 | 0.000 | -.12751 | .02758 | -4.62 | 0.000 |
| educat | -.00913 | .01680 | -0.54 | 0.587 | -.11120 | .01865 | -5.96 | 0.000 |
| sbp | .05066 | .00313 | 16.18 | 0.000 | .07892 | .00338 | 23.35 | 0.000 |
| vituse | -.72097 | .10752 | -6.71 | 0.000 | -.64256 | .10176 | -6.31 | 0.000 |
| smoker | -.47851 | .10456 | -4.58 | 0.000 | -.34981 | .10033 | -3.49 | 0.001 |
| _cons | 12.70443 | .49020 | 25.92 | 0.000 | 10.36452 | .52213 | 19.85 | 0.000 |

(B) 利用数据设计因素的分析

```
. svyset [pweight=wgt], strata(stra) psu(psu)
. svyregress bmi age agesq black hispanic pir educat sbp vituse smoker, deff

Survey linear regression
pweight:  wgt                                  Number of obs    =      9920
Strata:   stra                                 Number of strata =        23
PSU:      psu                                  Number of PSUs   =        46
                                               Population size  =   9920.06
                                               F(   9,      15) =     71.84
                                               Prob > F         =    0.0000
                                               R-squared        =    0.1380
```

| bmi | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] | Deff |
|---|---|---|---|---|---|---|---|
| age | .39778 | .02110 | 18.85 | 0.000 | .35412 | .44143 | 2.0539 |
| agesq | -.00421 | .00023 | -18.02 | 0.000 | -.00469 | -.00373 | 2.3647 |
| black | .96291 | .22418 | 4.30 | 0.000 | .49916 | 1.42666 | 1.5778 |
| hispanic | .64825 | .20430 | 3.17 | 0.004 | .22562 | 1.07087 | .8897 |
| pir | -.12751 | .05624 | -2.27 | 0.033 | -.24855 | -.01117 | 4.5323 |
| educat | -.11203 | .02703 | -4.11 | 0.000 | -.16712 | -.05529 | 2.1457 |
| sbp | .07892 | .00514 | 15.35 | 0.000 | .06828 | .08956 | 1.8798 |
| vituse | -.64256 | .17793 | -3.61 | 0.001 | -1.01063 | -.27449 | 3.0546 |
| smoker | -.34982 | .20405 | -1.71 | 0.100 | -.77192 | .07229 | 4.0343 |
| _cons | 10.36452 | .80124 | 12.94 | 0.000 | 8.70704 | 12.02201 | 2.3041 |

多元回归分析的结果。自变量与前面的描述性分析所使用的相同。此外，我们还包括了年龄平方，以控制可能存在的年龄对 BMI 的非线性效应。简单起见，这个例子中没有包含交互项，因为包含交互项除了会加深多重共线性问题外，还会增加 $R^2$。我们在这个分析中使用了缺失值的填补值。回

归系数几乎与没有使用填补值的分析结果相同，系数的标准误差也类似。

表格的顶部显示了不考虑设计特征时未加权和加权后的分析结果。regress 命令同时用于未加权和加权后分析，加权后分析中权重由(w = wgt)命令确定。首先，我们重点来看令人失望的很低的 $R^2$ 值，在未加权分析中是 0.12，在加权后分析中为 0.14。这表示大部分 BMI 的变异情况无法由模型解释。模型还包括了其他重要的变量。也许令人满意的预测 BMI 的模型设定无法用 NHANES III 数据实现。

未加权和加权后分析两者都显示年龄与 BMI 正相关，而年龄平方则与 BMI 负相关。这表示年龄的作用是曲线的，正如我们期望的，年龄越大趋势越弱。贫困指数和教育与 BMI 负相关。至于二元变量的回归系数，黑人和西班牙裔这两个变量的系数都是正的，表示这两个少数种族群体分别比他们的对立群体(非黑人、非西班牙裔)具有更大的 BMI。收缩压与 BMI 正相关，而使用维他命的人(他们可能更加关注自己的健康)则比不使用的人具有更低的 BMI。那些曾经抽烟的人与没有抽过烟的人相比，二者之间的 BMI 值差距不超过 0.5 个点。

未加权和加权后分析间存在着细小差别。虽然在未加权分析中，教育的作用很小(beta 系数 =－0.009)，但在加权后分析中其绝对值则变大了很多(beta 系数 =－0.111)。如果前期分析并没有使用抽样权重，我们就可能忽略了教育作为一个重要预测因素的作用。这个例子清楚显示了(本章开头所讨论的)利用 PPS 子样本进行前期分析的优势。抽烟状态的负系数值稍微减弱了，意味着相比其他人，抽烟对 BMI

的负面影响在被过抽样的人群中更加严重。这里再次说明了抽样权重的重要性。这个分析还指出了在前期分析中使用 PPS 子样本而不是未加权分析的优势。

把权重和设计特征同时考虑进来的分析结果列在表格底部。这个分析使用的是 svyregress 命令。回归系数的估计值和 $R^2$ 与加权后分析中的结果一样,因为在这个估计中我们使用的是相同的公式。但是,系数的标准误差和 $t$ 统计值则与加权后分析中的非常不一样。估计到的回归系数的设计效应在 0.89(Hispanics)和 4.53(pir,贫困与收入比)之间。我们再次看到相对 SRS 设计,复杂调查设计可能会对某些变量带来更大的方差,但并不是对所有的变量都这样。在这个特殊的例子中,前期分析中得到的一般的分析结果在最终的分析中都得到了支持,虽然对除了一个变量以外的所有其他变量来说回归系数的标准误差都增大了。

比较表 6.2 和表 6.4 中的设计效应,我们发现,回归系数的设计效应比均值和百分比的稍微小点。所以,把从均值和总体中估计到的设计效应应用到回归系数中(当数据中不包含集群信息时)会导致太保守的结论。如果回归模型控制了某些集群对集群的变异性,那么就可能在回归分析中得到更小的设计效应。比如,如果处于相同集群中的人具有类似 BMI 的部分原因是因为他们具有类似的年龄和教育程度,那么在回归模型中按年龄和教育程度进行调整就可能解释掉一部分集群对集群的变异性。这样集群效应对模型的残差的影响就可能更小。

回归分析也可以用 SUDAAN 的 REGRESS 程序执行,过程如下:

```
PROC REGRESS DESIGN = wr;
NEST stra psu;
WEIGHT wgt;
MODEL = bmi age agesq black hispanic pir educat sbp
vituse smoker;
RUN;
```

# 第4节 | 列联表分析

分析两个离散变量间相关性的最简单的方法是二维表格。如果数据来自 SRS,我们可以用 Pearson 卡方统计值检验关于独立的零假设。对基于复杂调查数据的二维列表分析,需要对这个检验方法进行修改以考虑调查设计的影响。已有多种不同的检验统计值被提出来。Koch、Freeman 和 Freeman(1975)提议使用 Wald 统计值[6],且已被广泛使用。Wald 统计值通常被转化成 $F$ 统计值以决定 $p$ 值。在 $F$ 统计值中,分子中的自由度与表格的维度联系在一起,分母中的自由度反映的是调查设计。后来,Rao 和 Scott(1984)提出了对数似然统计值的修正方法,使用具非整数自由度的 $F$ 统计值。基于一个模拟研究(Sribney, 1998), Stata 把 Rao-Scott 修正统计值设为默认程序,但 Wald 卡方和对数线性 Wald 统计值也是可用的选项。另一方面,SUDAAN 在它的 CROSSTAB 程序中用 Wald 统计值。在多数情况下,这些统计值得到的结论是相同的。

表 6.5 列出了用 Stata 进行的二维列表分析情况。这个分析列出了维他命使用与编码为三个类别的教育年限(1=少于 12 年,2=12 年,3=多于 12 年)之间的相关情况。在(A)部分,普通的卡方分析忽略了权重和数据结构的影响。

## 表 6.5 NHANES III 阶段 II 中按教育程度划分的美国成年人的维他命使用情况比较 ($n = 9920$)：用 Stata 进行的分析

(A)
```
. tab vituse edu, column chi
----------------------------------------------------------
           |                 edu
    vituse |         1          2          3 |     Total
-----------+---------------------------------+----------
         0 |      2840       1895       1372 |      6107
           |     68.43      61.89      50.66 |     61.56
-----------+---------------------------------+----------
         1 |      1310       1167       1336 |      3813
           |     31.57      38.11      49.34 |     38.44
-----------+---------------------------------+----------
     Total |      4150       3062       2708 |      9920
           |    100.00     100.00     100.00 |    100.00
          Pearson chi2(2) = 218.8510    Pr = 0.000
----------------------------------------------------------
```

(B)
```
. svyset [pweight=wgt], strata(stra) psu(psu)
. svytab vituse edu, column ci pearson wald

pweight:  wgt                                Number of obs      =      9920
Strata:   stra                               Number of strata   =        23
PSU:      psu                                Number of PSUs     =        46
                                             Population size    =   9920.06
-------------------------------------------------------------------------
          |                            edu
   vituse |              1              2              3          Total
----------+--------------------------------------------------------------
        0 |          .6659          .6018          .4834          .5705
          |  [.6307,.6993]  [.5646,.6379]  [.4432,.5237]  [.5452,.5955]
        1 |          .3341          .3982          .5166          .4295
          |  [.3007,.3693]  [.3621,.4354]  [.4763,.5568]  [.4045,.4548]
    Total |              1              1              1              1
-------------------------------------------------------------------------
  Key:  column proportions
        [95% confidence intervals for column proportions]
  Pearson:
    Uncorrected   chi2(2)           =  234.0988
    Design-based  F(1.63, 37.46)    =   30.2841      P = 0.0000
  Wald (Pearson):
    Unadjusted    chi2(2)           =   51.9947
    Adjusted      F(2, 22)          =   24.8670      P = 0.0000
-------------------------------------------------------------------------
```

(C)
```
. svyset [pweight=wgt], strata(stra) psu(psu)
. svytab vituse edu, subpop(hispanic) column ci wald

pweight:  wgt                                Number of obs      =      9920
Strata:   stra                               Number of strata   =        23
PSU:      psu                                Number of PSUs     =        46
                                             Population size    =   9920.06
Subpop.:  hispanic==1                        Subpop. no. of obs =      2593
                                             Subpop. size       =   539.043
-------------------------------------------------------------------------
          |                            edu
   vituse |              1              2              3          Total
----------+--------------------------------------------------------------
        0 |          .7382          .6728          .5593          .6915
          |  [.6928,.7791]  [.6309,.7122]  [.4852,.6309]  [.6509,.7293]
        1 |          .2618          .3272          .4407          .3085
          |  [.2209,.3072]  [.2878,.3691]  [.3691,.5148]  [.2707,.3491]
    Total |              1              1              1              1
-------------------------------------------------------------------------
  Key:  column proportions
        [95% confidence intervals for column proportions]
  Wald (Pearson):
    Unadjusted    chi2(2)           =   47.1625
    Adjusted      F(2, 22)          =   22.5560      P = 0.0000
-------------------------------------------------------------------------
```

教育与维他命使用之间在统计上存在着显著相关，教育程度越高的人越倾向于使用维他命。维他命使用者的百分比处于32%(在最低的教育水平中)到49%(在最高的教育水平中)之间。(B)部分列出了用相同的数据但把调查设计考虑进来时的分析结果。按教育程度划分，加权后维他命使用者的比例其变化情况比未加权时稍微大了一点，处于33%(第1级教育程度)到52%(第3级教育程度)之间。注意当使用了 ci 命令时，Stata 会对单元百分比计算置信区间。

这个分析要求同时报告 Pearson 和 Wald 卡方统计值。基于加权后频数的未经修正的 Pearson 卡方比(A)中的卡方值稍微大点，反映出加权后百分比的变化情况要大些。但是，基于未经修正的 Pearson 卡方统计值，我们无法估计出能反映复杂设计的合适的 $p$ 值。合适的 $p$ 值可以从基于设计的自由度为1.63和37.46的 $F$ 统计值30.28中评估出来，它以由 Rao-Scott 修正产生的检验程序为基础。未经调整的 Wald 卡方检验统计值为51.99，但 $p$ 值的产生必须以调整后的 $F$ 统计值为基础。两个 $F$ 统计值中的分母自由度都反映了抽样设计中 PSU 和阶层的数量。调整后的 $F$ 统计值只比 Rao-Scott 的 $F$ 统计值稍微小了一点。这些检验统计值中的任意一种都会得出相同的结论。

在(C)部分，我们对西班牙裔人口进行了子群体分析。在这个分析中我们使用了完整的数据文件。分析基于2593个观察值，但当考虑了抽样权重之后，只代表539个人。西班牙裔中维他命使用者的比例(31%)显著低于总体人口中维他命使用者的比例(43%)。正如调整后的 $F$ 统计值所示，

西班牙裔人口中，教育与维他命使用之间在统计上仍然存在显著相关。

现在我们来看看三维列表。利用 NHANES III 阶段 II 中的成年人样本数据，我们检验按教育程度划分的维他命使用的性别差异情况。这是一个 2×2×3 的表格，我们可以在每个教育程度中都进行一个二维列表分析。表 6.6 列出了用 SAS 和 SUDAAN 进行的三个 2×2 列表分析结果。表格顶部显示的是忽略了调查设计的分析结果。在最低的教育水平中，男性维他命使用者的百分比比女性的低，卡方统计

**表 6.6 NHANES III 阶段 II 中按教育程度划分的美国成年人的维他命使用性别差异分析 ($n = 9920$)：用 SAS 和 SUDAAN 进行的分析**

(A) 用 SAS 做的未加权的分析

```
proc freq;
tables edu*sex*vituse / nopercent nocol chisq measures cmh;
run;
        [Output summarized below]

Level of education:  Less than H.S.      H.S. graduate       Some college
Vitamin use status:  (n)     User        (n)     User        (n)     User

  Gender-   Male:    (1944)   26.34%     (1197)   31.91%     (1208)   43.54%
          Female:    (2206)   36.17      (1865)   42.09      (1500)   54.00

     Chi-square:      46.29               32.02               29.27
        P-value:      <.0001              <.0001              <.0001
     Odds ratio:      0.63                0.64                0.66
         95% CI:      (0.56, 0.72)        (0.56, 0.75)        (0.56, 0.76)

               CMH chi-square:        107.26     (p<.0001)
          CMH common odds ratio:        0.64     95%CI: (0.59, 0.70)
                              ------------------------------------------
```

(B) 用 SUDANN 做的调查分析

```
proc crosstab design=wr;
nest stra psu;
weight wgt;
subgroup edu sex vituse;
levels     3   2    2;
tables edu*sex*vituse;
print nsum wsum rowper cor upcor lowcor chisq chisqp cmh cmhpval;
run;
        [Output summarized below]

Level of education:  Less than H.S.      H.S. graduate       Some college
Vitamin use status:  (n)*    User        (n)*    User        (n)*    User

   Gender-  Male:    (1274.9)  28.04%    (1432.2)  32.66%    (2031.1)  45.62%
          Female:    (1299.7)  38.62     (1879.4)  45.43     (2002.7)  57.37

     Chi-square:      19.02               38.01               10.99

        P-value:      .0002               <.0001              .0030
     Odds ratio:      0.62                0.58                0.62
         95% CI:      (0.50, 0.77)        (0.49, 0.69)        (0.48, 0.80)

               CMH chi-square:        42.55     (p<.0001)
*Weighted sum
```

值显示这个差别在统计上是显著的。另一个检验 2×2 列表中的相关性的方法是计算比数比。

在这个表格中,男性维他命使用情况的比数为 0.2634/(1－0.2634)＝0.358,女性的则为 0.3617/(1－0.3617)＝0.567。男性比数与女性比数的比为 0.358/0.567＝0.63,意味着男性使用维他命的比数是女性使用维他命比数的 63%。95%的置信区间内不包含 1,表示这种差别在统计上是显著的。在三个不同程度的教育水平间,比数比结果是一致的。由于这些比数比是一致的,因此我们可以把这三个不同教育程度的 2×2 列表合起来。然后就可计算 Cochran-Mantel-Haenszel(CMH)卡方($df=1$)和 CMH 普通比。按教育调整后的比数比为 0.64,其 95%置信区间不包含 1。

表 6.6 下半部分列出了考虑调查设计后使用 SUDAAN 中的 CROSSTAB 程序进行相同分析的结果。在 PROC 命令句式中,DESIGN＝wr 指定有放回抽样,表示并没有使用有限总体校正。NEST 命令指定阶层和 PSU 变量,WEIGHT 命令给出了权重变量。SUBGROUP 命令指出了 3 个离散变量,而 LEVELS 命令则明确了每个离散变量的(取值)类别的数量。TABLES 命令界定了列联表的形式。PRINT 命令要求报告 nsum(频数)、wsum(加权后频数)、rowper(行百分比)、cor(粗比率比)、upcor(cor 的上限)、lowcor(cor 的下限)、chisq(卡方统计值)、chisqp(卡方统计值的 $p$ 值)、cmh(CMH 统计值)、cmhpval(CMH 的 $p$ 值)。

维他命使用的加权后百分比与未加权百分比稍微不同。在表格上半部的三个不同的分析中,除了在中间那个教育水平之外,Wald 卡方值比 Pearson 卡方值小。虽然在更低的教

育水平中比数比几乎保持不变，但在中间及更高的教育水平中却降低了。SUDAAN 中的 CROSSTAB 程序没有计算普通比数比，但可从 logistic 回归分析中得到，这将在下文中讨论。

# 第5节 | logistic 回归分析

前面讲到的线性回归分析对很多社会科学家来说可能用处不大,因为在社会科学研究中很多变量通常都用类别(名义的或者次序的)进行测量。有一系列用于分析分类数据的统计方法,从前一节所提到的基本的跨列表分析到具各种不同联结函数的广义线性模型等等。正如 Knoke 和 Burke (1980)所观察到的,建模的方法在社会科学研究中改革了列联表分析方法,废除了大部分用于研究离散类别测量的变量之间的关系的旧方法。有两种方法被社会科学家广泛应用:使用最大似然方法的对数线性模型(Knoke & Burke, 1980; Swafford, 1980)和加权后最小二乘法(Forthofer 7 Lehnen, 1981; Grizzled、Starmer & Koch, 1969)。这两种用于分析复杂调查数据的方法在本书的第一版中有所阐述。在这些模型中,单元比例或者它们的函数(如对数线性模型中的自然对数)用列联表中的(各种因素的)效应的线性组合表示。由于这些模型受限于列联表,因此在分析中无法包含连续的自变量。

在过去的十年间,社会科学家已经开始更加频繁地使用 logistic 回归分析,因为它可以把更多的解释性变量涵盖进来,其中就包括连续变量(Aldrich & Nelson, 1984; DeMaris,

1992；Hosmer & Lemeshow，1989；Liao，1994）。Logistic 回归和其他具不同联结函数的广义线性模型现在已经嵌入到复杂调查数据分析的(统计)软件包里了。调查分析者可以从一系列模型中选择最合适的模型。用 Stata 和 SUDAAN，logit 模型的应用阐释如下。

方程 6.1 代表普通线性回归分析，检验的是一个连续因变量与一个或多个自变量之间的关系。logistic 回归是一种用于检验一个类别因变量与一系列自变量之间关系的方法。下面这个方程表示对一个二分类结果变量 $Y$ 建立模型的方式，$i = 1, 2, \cdots, n$：

$$\log[\pi_i/1-\pi_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_{p-1} x_{p-1,i} \quad [6.2]$$

在方程 6.2 中，$\pi_i$ 是 $y_i = 1$ 的概率。这是一个具对数比或 logit 联结函数的广义线性模型。这里用最大似然的方法(Eliason，1993)来估计参数，而不是最小二乘估计法。因为这里要解的联立方程对参数并不是线性的，所以使用的是迭代技术。最大似然理论同时还可估计这些参数($\beta$)的估计值的协方差矩阵，前提是个体观察值是随机且相互独立的。

正如方差模型的分析一样，如果一个变量有 $l$ 个类别，那么我们在模型中只用 $l-1$ 个类别。我们将从被忽略类别(或者参照组)的效应中测量这 $l-1$ 个类别的效应。$\beta$ 的估计值($\hat{\beta}$)是模型中对应类别与被忽略类别之间的 logit 的差别，即模型中对应类别与被忽略类别的比数比的自然对数。因此，$e^{\hat{\beta}}$ 就等于对模型中的其他变量进行调整后得到的比数比。logistic 回归结果通常以比数比的方式总结和解释(Liao，

1994:第 3 章)。

在复杂调查数据中,需要对最大似然估计进行调整,因为每个观察值都有一个抽样权重。结合了权重的最大似然解通常被称为 pseudo 或者加权后最大似然估计(Chambless & Boyle, 1985; Roberts、Rao & Kumar, 1987)。虽然点估计是用 pseudo 似然方法计算的,但 $\hat{\beta}$ 的协方差矩阵却是用第 4 章讨论的其中一种方法估计的。正如前面提到的,与这个协方差矩阵相关的自由度大约等于 PSU 的数量减去阶层的数量。因此,标准的模型拟合的似然比检验不能用于调查数据的 logistic 回归分析之中。相反要使用调整后的 Wald 检验。

在 logistic 回归模型中,对合适预测变量的选择或纳入与线性回归中的过程类似。当分析的是一个大型调查数据库时,前一节所描述的前期分析策略在为 logistic 回归分析做准备时非常有用。

为阐释 logistic 回归分析,我们用 Stata 对表 6. 6 用到的数据进行分析。分析结果见表 6. 7。我们稍微编辑了一下 Stata 输出结果以适应表格大小。结果变量是维他命使用,解释变量包括性别(1 = 男性;0 = 女性)和教育程度。这个模型基于表 6. 6 所显示的 CMH 统计值,没有包含交互项。首先,我们进行了标准的 logistic 回归分析,忽略了权重和设计特征。结果列于表格(A)部分。在 logit 命令之前加上 xi 并在相应的变量名前面加上 i. ,Stata 会自动对离散变量执行效果或二分类编码。

表中的输出结果列出了每个离散变量被忽略的类别。在这个例子中,“男性”这个类别,在模型中它的效应是从参

**表 6.7 NHANES III 阶段 II 中美国成年人性别和教育程度对维他命使用影响的 logistic 回归分析 ($n = 9920$)：用 Stata 进行的分析**

**(A) 标准 logistic 回归（忽略设计效应和权重）**

```
. xi: logit vituse i.male i.edu

       i.male            _Imale_0-1          (naturally coded; _Imale_0 omitted)
       i.edu             _Iedu_1-3           (naturally coded; _Iedu_1 omitted)
       Iteration 0:   log likelihood = -6608.3602
       Iteration 1:   log likelihood = -6445.8981
       Iteration 2:   log likelihood = -6445.544
       Iteration 3:   log likelihood = -6445.544
Logit estimates                                    Number of obs   =      9920
                                                   LR chi2(3)      =    325.63
                                                   Prob > chi2     =    0.0000
Log likelihood = -6445.544                         Pseudo R2       =    0.0246
------------------------------------------------------------------------------------
    vituse |   Coef.  Std. Err.     z    P>|z|  [95% Conf.Int.] Odds Ratio [95% Conf.Int.]
-----------+------------------------------------------------------------------------
  _Imale_1 |  -.4418   .0427   -10.34   0.000  -.5256  -.3580    .6429     .5912  .6990
   _Iedu_2 |   .2580   .0503     5.12   0.000   .1593   .3566   1.2943    1.1773 1.4285
   _Iedu_3 |   .7459   .0512    14.56   0.000   .6455   .8462   2.1082    1.9069 2.3308
     _cons |  -.5759   .0382   -15.07   0.000  -.6508  -.5010
------------------------------------------------------------------------------------
```

**(B) 拟合优度检验**

```
. lfit

Logistic model for vit, goodness-of-fit test
      number of observations =      9920
 number of covariate patterns =         6
             Pearson chi2(2) =      0.16
                 Prob > chi2 =    0.9246
```

**(C) 利用权重和设计因素的调查 logistic 回归**

```
. svyset [pweight=wgt], strata (stra) psu(psu)
. xi: svylogit vituse i.male i.edu

       i.male            _Imale_0-1          (naturally coded; _Imale_0 omitted)
       i.edu             _Iedu_1-3           (naturally coded; _Iedu_1 omitted)

Survey logistic regression
pweight:  wgt                                      Number of obs    =      9920
Strata:   stra                                     Number of strata =        23
PSU:      psu                                      Number of PSUs   =        46
                                                   Population size  =   9920.06
                                                   F(   3,     21)  =     63.61
                                                   Prob > F         =    0.0000
----------------------------------------------------------------------------------------------
    vituse |   Coef.  Std. Err.     t    P>|t|  [95% Conf. Int.]    Deff   Odds Rat.[95% Conf. Int.]
-----------+----------------------------------------------------------------------------------
  _Imale_1 |  -.4998   .0584    -8.56   0.000  -.6206  -.3791   1.9655    .6066    .5376   .6845
   _Iedu_2 |   .2497   .0864     2.89   0.008   .0710   .4283   2.4531   1.2836   1.0736  1.5347
   _Iedu_3 |   .7724   .0888     8.69   0.000   .5885   .9562   2.8431   2.1649   1.8013  2.6019
     _cons |  -.4527   .0773    -5.86   0.000  -.6126  -.2929   2.8257
----------------------------------------------------------------------------------------------
```

**(D) 检验系数的线性合并**

```
. lincom _Imale_1+_Iedu_3, or

 ( 1)  _Imale_1 + _Iedu_3 = 0.0
------------------------------------------------------------------------
     vituse | Odds Ratio  Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+-----------------------------------------------------------
        (1) |   1.3132     .1518     2.36   0.027      1.0340     1.6681
------------------------------------------------------------------------
```

照组“女性”的效应中测量出来的。至于教育程度，低于高中毕业是参照类别。似然比卡方值等于 325.63（自由度 $df = 3$），$p$ 值为 $< 0.00001$，我们否定性别与教育程度这两种因素一起对维他命使用没有影响这个假设，表示存在着显著影响。但 pseudo $R^2$ 表示维他命使用的大部分变异情况不

能被这两个变量所解释。性别和教育的参数估计值、它们的标准误差估计值以及相应的检验统计值也列于表中,所有的因素都显著。

在模型命令中加入 or 会得到比数比,而不是 $\beta$ 系数。估计到的男性的比数为 0.64,表示在对教育进行了调整之后,男性在使用维他命方面的比数[①]是女性比数的 64%。这个比数比与表 6.6 中的 CMH 普通比数比相同。比数比的显著性可以用 $z$ 检验或者置信区间进行检验。教育程度的第三个类别的比数比表示,具大学教育程度的人其使用维他命的可能性是那些具相同性别但教育年限少于 12 的人的两倍。没有任何一个置信区间包含 1,表示所有的影响都是显著的。

表 6.7 的(B)部分列出了拟合优度统计值(自由度 $df=2$ 的卡方值)。$p$ 值很大表示主效应模型适合数据情况(与饱和模型没有显著差别)。

在这种简单的情况下,与模型拟合优度相关的 2 个自由度也可以解释为与性别—教育交互项相关的 2 个自由度。因此,关于维他命的使用比例,并不存在性别与教育的交互作用,这也证实了表 1.15 中的 CMH 分析结果。

表 6.7 的(C)部分显示了把调查设计考虑在内时,对相同数据进行 logistic 回归分析的结果。对数似然值没有列出来,原因在于使用的是 pseudo 似然法。这里我们用的是 $F$ 统计值而不是似然比统计值。同样,$p$ 值表示主效应模型是对零模型的一个显著的改进。由于抽样权重的原因,估计到的参数和比数比有变化。正如在设计效应中反映出来的一

---

① 即使用的概率/不使用的概率。——译者注

样，$\beta$系数的标准误差增加了。虽然标准误差增加了，但性别和教育的$\beta$系数显著地不等于0。对教育调整后的男性的比数从0.64下降到0.61。虽然对第二个教育水平保持不变，但$p$值却显著变大了，从<0.0001上升到0.008，原因在于我们把设计（因素）也考虑在内了。

在运行了logistic回归之后，(D)部分显示了对参数线性组合的效应进行检验的结果。我们想要检验男性的参数与第三个教育水平的参数之和等于0这个假设。由于没有交互效应，因此得到的1.3这个比数比可以解释为，具大学教育的男性其使用维他命的比数相对于参照组（少于12年教育的女性）的比数高30%。SUDAAN也用于执行logistic回归分析，在单机版中使用LOGISTIC程序或者在SAS的可调用版本（不同的用以与SAS中的标准logistic程序区分开来的名字）中使用RLOGIST程序。

最后，logistic回归模型还可对综合估计建立预测模型。由于大部分健康调查是为估计全国性的统计值而设计的，因此我们很难对小区域进行健康特征的估计。得到小地区估计值的一种方法就是利用全国性的健康调查和地区性的人口信息进行综合估计。LaVange、Lafata、Koch和Shah(1996)利用一个拟合了NHIS（美国国家健康访问调查）和ARF（地区资源档案）的logistic回归模型对美国各州各县老人活动受限的普遍情况进行估计。由于NHIS以复杂调查设计为基础，他们用SUDAAN对NHIS的活动受限预测因素以及ARF中的县级变量进行拟合，建立了logistic回归模型。然后，以模型为基础的概率估计值被用于计算小地区的活动受限估计值。

# 第6节 其他 logistic 回归模型

上面所讨论的二分类 logistic 回归模型可被扩展用以处理多于两个回应类别的情况。这种回应类别中有些是次序的,如感知的健康状况:很好、好、一般和差。有些可能是名义上的,比如在宗教崇拜中。这些次序的和名义上的结果变量可被当做是一系列离散和连续自变量的函数来进行检验。利用 Stata 或者 SUDAAN,这种建模方法可用于复杂调查数据。在这一部分,我们举了两个这种分析的例子,但没有对它们进行深入讨论和分析。关于这种方法的详细描述及分析请见 Liao(1994)。

为阐释次序 logistic 回归模型,我们检验了基于 BMI 的肥胖类别。公共卫生营养学家们用下面的标准把 BMI 分成几种肥胖程度:极度肥胖 (BMI $\geqslant$ 30)、超重(25 $\leqslant$ BMI $<$ 30)、正常(18.5 $\leqslant$ BMI $<$ 25)、偏瘦(BMI $<$ 18.5)。基于 NHANES III 阶段 II 的数据,18%的美国成年人为极度肥胖、34%超重、45%正常、3%偏瘦。我们想要检验这四种肥胖程度(bmi2:1=极度肥胖、2=超重、3=正常、4=偏瘦)与一系列解释变量包括年龄(连续变量)、教育、黑人和西班牙裔等之间的关系。

对于这四个肥胖的有序类别,我们对以下三组概率建模

使之成为解释变量的函数：

Pr{极度肥胖}相对于 Pr{其他所有类别}

Pr{极度肥胖＋超重}相对于 Pr{正常＋偏瘦}

Pr{极度肥胖＋超重＋正常}相对于 Pr{偏瘦}

然后三个二分类 logistic 回归模型可分别用于上述每一组比较。但是，鉴于这个肥胖类别的自然排序，我们可以在成比例发生比假设的基础上，同时考虑这三个二分类模型来估计解释变量的“平均”效应。这里假设的是，每个不同结果类别的回归线之间是相互平行的，但截距可以不同（这个假设需要用卡方统计值进行检验，检验结果没有列于表中）。下面这个方程代表了 $j=1, 2, \cdots, c-1$（$c$ 是因变量的类别数量）的模型：

$$\log\left(\frac{\Pr(category \leqslant j)}{\Pr[category \geqslant (j+1)]}\right)=\alpha_j+\sum_{i=1}^{p}\beta_i x_i \quad [6.3]$$

从这个模型，我们估计 $(c-1)$ 个截距项和一系列 $\hat{\beta}$。

表 6.8 显示了以上用 SUDAAN 进行的分析结果。SUDAAN 命令列于表格顶部。第一个命令 PROC MULTILOG 明确了所用的程序。DESIGN，NEST 和 WEIGHT 的具体说明与表 6.6 的一样。REFLEVEL 说明教育的第一个类别作为参照组（如果不明确指出则变量的最后一个类别作为参照组）。

分类变量列在 SUBGROUP 命令之后，这些变量每个所包含类别的数量列在 LEVELS 命令之后。MODEL 这一行命令列出了因变量，随后是一系列自变量。MODEL 命令中的关键词 CUMLOGIT 指明用的是成比例发生比模型。若没有这个关键词，SUDAAN 将会拟合多类别logistic 回归模

**表 6.8 NHANES III 阶段 II 中美国成年人教育、年龄和种族对肥胖程度影响的次序 logistic 回归分析($n = 9920$)：用 SUDAAN 进行的分析**

```
proc multilog design=wr;
nest stra psu;
weight wgt;
reflevel edu=1;
subgroup bmi2 edu;
levels   4    3;
model bmi2=age edu black hispanic/ cumlogit;
setenv decwidth=5;
run;

Independence parameters have converged in 4 iterations
-2*Normalized Log-Likelihood with Intercepts Only: 21125.58
-2*Normalized Log-Likelihood Full Model          : 20791.73
Approximate Chi-Square (-2*Log-L Ratio)          :   333.86
Degrees of Freedom                               :        5

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Cumulative Logit
Response variable: BMI2
-------------------------------------------------------------------
BMI2 (cum-logit),
  Independent                                                P-value
  Variables and          Beta                                T-Test
  Effects                Coeff.       SE Beta   T-Test B=0   B=0
-------------------------------------------------------------------
BMI2 (cum-logit)
  Intercept 1          -2.27467       0.11649    -19.52721   0.00000
  Intercept 2          -0.62169       0.10851     -5.72914   0.00001
  Intercept 3           2.85489       0.11598     24.61634   0.00000
AGE                     0.01500       0.00150      9.98780   0.00000
EDU
  1                     0.00000       0.00000          .         .
  2                     0.15904       0.10206      1.55836   0.13280
  3                    -0.20020       0.09437     -2.12143   0.04488
BLACK                   0.49696       0.08333      5.96393   0.00000
HISPANIC                0.55709       0.06771      8.22744   0.00000
-------------------------------------------------------------------
       -----------------------------------------------------
       Contrast             Degrees of               P-value
                            Freedom       Wald F     Wald F
       -----------------------------------------------------
       OVERALL MODEL          8.00000    377.97992   0.00000
       MODEL MINUS
         INTERCEPT            5.00000     36.82064   0.00000
       AGE                    1.00000     99.75615   0.00000
       EDU                    2.00000     11.13045   0.00042
       BLACK                  1.00000     35.56845   0.00000
       HISPANIC               1.00000     67.69069   0.00000
       -----------------------------------------------------
-----------------------------------------------------------
BMI2 (cum-logit),
  Independent
  Variables and                     Lower 95%     Upper 95%
  Effects              Odds Ratio   Limit OR      Limit OR
-----------------------------------------------------------
AGE                       1.01511     1.01196       1.01827
EDU
  1                       1.00000     1.00000       1.00000
  2                       1.17239     0.94925       1.44798
  3                       0.81857     0.67340       0.99503
BLACK                     1.64372     1.38346       1.95295
HISPANIC                  1.74559     1.51743       2.00805
-----------------------------------------------------------
```

型，这个模型会在下文进行讨论。最后，SETENV 命令要求在发表结果时报告五个小数点。

输出结果显示了三个截距项和自变量的一组系数估计结果。第二个表框中的结果表明所有主效应都显著。第三个表框中的比率比可以用与二分类logistic 回归中相同的方式进行解释。在控制了其他自变量之后，西班牙裔比非西班牙裔具有 1.7 倍甚至更高的属于极度肥胖的比率。在解释这些结果之前，我们必须检查是否满足了成比例发生比假设，但是输出结果并没有给出任何检验这个假设的统计值。为检验这个假设，我们运行了三个普通 logistic 回归分析（极度肥胖相对所有其他，极度肥胖＋超重相对正常＋偏瘦，极度肥胖＋超重＋正常相对偏瘦）。年龄对应的三个比率比分别为 1.005、1.012 和 1.002，它们与表 6.8 底部所显示的 1.015 很接近。其他自变量对应的比率比也相当类似。我们认为成比例发生比假设看来可以接受。

利用 svyolog 程序，Stata 还可用于拟合成比例发生比模型，但它所拟合的模型稍微有所不同。在方程 6.3 中，一组 $\beta_i x_i$ 项加在截距项之后，但在 Stata 模型中它被去掉了。因此，从 Stata 中估计到的 $\beta$ 系数其符号与从 SUDAAN 中得到的相反，而绝对值则相同。这意味着从 Stata 中得到的比数比是从 SUDAAN 中得到的比数比的相反数。这两种方法给出的截距项估计值是完全相同的。Stata 用的术语是 cut 而不是 intercept。

对名义的结果类别，可以使用多类别 logistic 回归模型。用这个模型，我们可以检验一个多类别名义结果变量（没有顺序的）与一组解释变量之间的关系。这个模型指定结果变

量中的某一个类别为基准类别,并估计相对于这个基准变量,处于第 $j$ 个类别的概率的比数的对数。这个比数称为相对风险,而这个比数的对数则称为广义 logit。

我们使用了上面用到的相同的肥胖程度。虽然我们明显地看到肥胖程度具有顺序,我们在此把它视为名义变量,因为我们想拿其他肥胖程度与正常这个类别进行比较。相应地,我们对肥胖程度的编码做了修改(bmi3:1=极度肥胖、2=超重、3=偏瘦、4=正常[基准类别])。我们使用了三个预测变量,包括年龄(连续变量)、性别(1=男性[参照组]、2=女性)和目前的吸烟状态(1=现在抽烟、2=从未抽过[参照组]、3=以前抽过现在不抽)。下面的方程代表了这个模型:

$$\log\left(\frac{\Pr(obese)}{\Pr(normal)}\right)=\beta_{0,1}+\beta_{1,1}(age)+\beta_{2,1}(male)+\beta_{3,1}(c.smoker)+\beta_{4,1}(p.smoker)$$

$$\log\left(\frac{\Pr(overweight)}{\Pr(normal)}\right)=\beta_{0,2}+\beta_{1,2}(age)+\beta_{2,2}(male)+\beta_{3,2}(c.smoker)+\beta_{4,2}(p.smoker)$$

$$\log\left(\frac{\Pr(underweight)}{\Pr(normal)}\right)=\beta_{0,3}+\beta_{1,3}(age)+\beta_{2,3}(male)+\beta_{3,3}(c.smoker)+\beta_{4,3}(p.smoker)$$

[6.4]

我们使用 SUDAAN 来拟合上面的模型,结果列于表 6.9 中(我们稍微编辑了一下 Stata 输出结果以适应表格大小)。SUDAAN 中所有的命令与前面成比例发生比模型中的命令类似,除了在 MODEL 这一行命令中省略了 CUMLOGIT。Stata 中的 svymlogit 程序也可用于拟合多类别回归模型。

**表 6.9 NHANES III 阶段 II 中美国成年人性别和抽烟状态对肥胖程度影响的多类别 logistic 回归分析 ($n = 9920$)：用 SUDAAN 进行的分析**

```
proc multilog design=wr;
nest stra psu;
weight wgt;
reflevel csmok=2 sex=2;
subgroup bmi2 csmok sex;
levels   4    3    2;
model bmi3=age sex csmok;
setenv decwidth=5;
run;

Independence parameters have converged in 6 iterations
Approximate ChiSquare (-2*Log-L Ration)         :  587.42
Degrees of Freedom                              :      12

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Generalized Logit
Response variable: BMI3
```

| BMI3 log-odds) | | Independent Variables and Effects | | | | |
|---|---|---|---|---|---|---|
| | | Intercept | AGE | SEX = 1 | CSMOK = 1 | CSMOK = 3 |
| 1 vs 4 | Beta Coeff. | -1.33334 | 0.01380 | 0.08788 | -0.39015 | -0.27372 |
| | SE Beta | 0.14439 | 0.00214 | 0.12509 | 0.07203 | 0.13206 |
| | T-Test B=0 | -9.23436 | 6.43935 | 0.70251 | -5.41617 | 2.07277 |
| | P-value | 0.00000 | 0.00000 | 0.48941 | 0.00002 | 0.04958 |
| 2 vs 4 | Beta Coeff. | -1.25883 | 0.01527 | 0.76668 | -0.24271 | -0.02006 |
| | SE Beta | 0.13437 | 0.00200 | 0.08275 | 0.11067 | 0.09403 |
| | T-Test B=0 | -9.36835 | 7.64830 | 9.26512 | -2.19307 | -0.21335 |
| | P-value | 0.00000 | 0.00000 | 0.00000 | 0.03868 | 0.83293 |
| 3 vs 4 | Beta Coeff. | -2.07305 | -0.01090 | -1.16777 | 0.33434 | 0.04694 |
| | SE Beta | 0.48136 | 0.00742 | 0.25280 | 0.30495 | 0.26168 |
| | T-Test B=0 | -4.30663 | -1.46824 | -4.61937 | 1.09637 | 0.17936 |
| | P-value | 0.00026 | 0.15558 | 0.00012 | 0.28427 | 0.85923 |

| Contrast | Degrees of Freedom | Wald F | P-value Wald F |
|---|---|---|---|
| OVERALL MODEL | 15.00000 | 191.94379 | 0.00000 |
| MODEL MINUS INTERCEP | 12.00000 | 68.70758 | 0.00000 |
| INTERCEPT | . | . | . |
| AGE | 3.00000 | 22.97518 | 0.00000 |
| SEX | 3.00000 | 64.83438 | 0.00000 |
| CSMOK | 6.00000 | 6.08630 | 0.00063 |

| BMI3(log-odds) | | Independent Variables and Effects | | | | |
|---|---|---|---|---|---|---|
| | | Intercept | AGE | SEX = 1 | CSMOK = 1 | CSMOK = 3 |
| 1 vs 4 | Odds Ratio | 0.26360 | 1.01390 | 1.09186 | 0.67695 | 0.76054 |
| | Lower 95% Limit | 0.19553 | 1.00941 | 0.84291 | 0.58323 | 0.57874 |
| | Upper 95% Limit | 0.35535 | 1.01840 | 1.41433 | 0.78573 | 0.99946 |
| 2 vs 4 | Odds Ratio | 0.28399 | 1.01539 | 2.15260 | 0.78450 | 0.98014 |
| | Lower 95% Limit | 0.21507 | 1.01120 | 1.81393 | 0.62397 | 0.80689 |
| | Upper 95% Limit | 0.37499 | 1.01959 | 2.55450 | 0.98633 | 1.19059 |
| 3 vs 4 | Odds Ratio | 0.12580 | 0.98916 | 0.31106 | 1.39702 | 1.04805 |
| | Lower 95% Limit | 0.04648 | 0.97408 | 0.18439 | 0.74341 | 0.60994 |
| | Upper 95% Limit | 0.34052 | 1.00447 | 0.52476 | 2.62526 | 1.80086 |

表 6.9 同时显示了 $\beta$ 系数和相对风险比率(标记为比数比)，同时也列出了标准误差和检验 $\beta = 0$ 的 $p$ 值。年龄在极度肥胖相对正常和超重相对正常的比较中都是一个显著的因素，但在偏瘦相对正常的比较中则不是。虽然性别在极度肥胖相对正常的比较中没有影响，但在另两个比较中却有。

我们来看比数比的表格，相对于正常这个类别男性处于超重的相对风险比率是女性的两倍以上，给定年龄和抽烟状态相同的情况下。相对于正常这个类别，在保持年龄和性别不变的情况下，现在抽烟的人其处于极度肥胖的相对风险比率只有那些从不抽烟的人的 68%。①

现有的软件也支持其他可用于分析复杂调查数据的统计模型。比如，SUDAAN 支持用于生存分析的 Cox 的回归模型(比例风险模型)，虽然截面调查很少提供纵向数据。利用 SUDAAN、Stata 和其他软件支持的程序，其他由不同的联结函数定义的广义线性模型也可应用到复杂调查数据中。

① 原文中为 0.68%，实为有误。——译者注

# 第7节 | 基于设计和基于模型的分析

到目前为止,所有的分析都依赖于基于设计的方法,因为分析中结合了抽样权重和设计特征。在把这些分析与基于模型的方法联系起来之前,我们先简单看看用于这些分析的调查数据。NHANES III 一共产生 2812 个 PSU,覆盖整个美国。这些 PSU 有时包括单个县,有时包括两个或以上相邻的县。这些县是行政单位,而调查并不是为了对这些单位分别进行估计而设计的。从这些单位中,81 个 PSU 被选中,选择概率与 PSU 的规模成比例——13 个来自确定的阶层。此外,从根据人口统计学特征而非地理位置形成的 34 个阶层中每层抽出两个 PSU。同样地,阶层并不是以定义人口参数为目的而设计的。抽样的第二个阶段涉及地区图块,包括城市或者城郊街区或者其他邻近的地理区域。具有更多的少数种族人口的图块用更高的概率进行选取。抽样的第三个阶段涉及把抽中的地区图块里面的所有家庭罗列出来,然后按照以图块特征为基础的比率对它们进行抽样。抽样的第四个阶段是对抽中的家庭里面的个人进行抽样以进行访问。这些次级单位用以辅助抽样而不是用于界定人口参数。公开的数据档案只包括阶层和 PSU,支持次级抽样单位的标识并不包括在内。抽样权重的计算基于被访问个人

的选择概率,此外还包括对无应答和事后分层的加权调整。前几章所讨论的许多分析问题是由大型社会和健康调查的执行方式造成的。现有的数据无法支持分层线性模型以融合多阶段选择设计。

由于受访个体的不等选择概率加上无应答和事后分层调整,若在 NHANES III 数据的描述性分析中忽略抽样权重,就会难以令人信服。在描述性分析中进行加权的原因已经很清楚。正如表 6.2 显示的,在未加权估计中年龄和与种族相关的变量的偏差很大。对所有的变量,加权后估计值的标准误差与未加权时的结果很类似,这意味着抽样权重的变异情况并不会增加方差。但是,当把 PSU 和阶层考虑进来时,加权后估计值的标准误差相当大,正如设计效应所反映的。降低方差的一种方法是使用包含辅助信息的模型。典型的例子是熟悉的比率和回归估计(Cochran, 1977:第 6 章)。在这些估计中,辅助信息采取的形式是已知的与目标变量相关的伴随变量的人口均值。辅助信息的使用可以扩展为分布函数的估计(Rao、Kovar & Mantel, 1990),但辅助信息在例行描述性分析中的使用是有限的,因为很难找到合适的辅助信息。

回归分析中模型的使用是显而易见的。表 6.4 中的未加权估计严格地以基于模型的方法为基础。它忽略了抽样权重和设计特征,但许多相关的设计信息已经包含在自变量之中:比如年龄(对年老者的过抽样)以及黑人和西班牙裔(对少数人口的过抽样)。其实,基于模型的系数估计值与加权后估计值类似,意味着基于模型的分析在这个情况下很合理。一个显著的例外是教育的系数,它在两种估计方法间非

常不同。在基于模型的分析中教育非常不显著,但在加权后分析中却非常显著。教育的效应无法被基于模型的分析检测到,因为对年长者而言,教育的作用在逐渐消失。虽然年龄已包含在模型中,但年龄和教育的交互作用却没有包含在模型中。这个例子说明,抽样权重的使用避免了模型的错误设定。

Korn 和 Graubard(1995b)利用 1988 年美国全国母婴健康调查数据,更进一步地阐释了在回归分析中使用抽样权重的好处。这个调查对低出生体重的婴儿进行了过抽样。估计到的胎龄对出生体重影响的回归线在未加权分析和加权后分析间非常不同。虽然未加权拟合均等反映了样本观察值且不对总体进行描述,但加权拟合把回归线拉到了估计到的总体所处的位置。这两个变量之间的关系其实是曲线的。相反,如果我们使用二次回归,那么未加权和加权后回归就会更具一致性。

正如上面对表 6.4 的分析结果所进行的讨论,对加权后和未加权回归间的差别进行细致的检验有时可以找到应该加入到方程中的重要变量或者交互项。未加权和加权后估计之间的差异表明,把抽样设计考虑在内可以避免可能出现的总体模型的错误设定。已有文献已经提出了几种用于检验加权后和未加权估计之间差异的统计值(DuMouchel & Duncan, 1983; Fuller, 1984; Nordberg, 1989)。Korn 和 Graubard(1995a)利用基于设计的变异性把这些检验统计值应用到 NHANES I 和 II 的数据中。他们建议当低效性很小时,使用基于设计的分析。否则,附加的建模假设可以融入到分析中去。他们已经注意到次级抽样单位并不对公众开

放这个问题,也已指出需要在大型健康调查中增加 PSU 的数量。这些检验受限于点估计,因此他们的结论不能应用到所有的情况中。Pfeffermann(1993、1996)提供了更详细的关于这些问题和与之相关问题的讨论。

基于设计的分析避免可能的模型错误设定这个事实,说明利用 SUDAAN、Stata 和其他复杂调查分析软件的分析适合 NHANES 数据。甚至在基于设计的分析中,我们也使用了模型来具体说明所感兴趣的参数,但推论把抽样权重考虑在内了。这种情况下,基于设计的分析可称为模型辅助方法(Sarndal、Swensson & Wretman, 1992)。基于设计的理论依赖于大的抽样规模来对参数做推论。对小样本来说基于模型的分析可能是更好的选择。在数据收集中没有使用概率抽样时,不存在应用基于设计的推论的基础。当实质理论和先前经验调查支持所提议的模型时,基于模型的方法可能更有道理。

相比在回归分析中,基于模型的分析的构思在列联表分析中更不易理解。我们已经讨论了把抽样方案考虑在内时基于设计的分析的基本原理。正如在回归分析中一样,很有必要关注加权后比例与未加权比例之间的差别。如果存在显著差别,我们就应该检查为什么会存在差别。在表 6.6 中,未加权比例和加权后比例很接近,但对高中毕业生和具大学教育的人来说,维他命使用和性别的加权后比数比比未加权比数比稍微小一些,而对教育程度低于高中毕业的人来说,加权后和未加权比数比则差不多相同。两个较高教育水平的轻微差别可能来自种族或其他因素。如果未加权和加权后比数比的差别大很多且归因于种族,我们就应该对不同

的种族群体分别检验这种相关性。在列联表中考虑附加因素的作用可以通过使用 logistic 回归模型实现。

logistic 回归中模型的使用以及相关问题与线性回归中的一模一样。对加权后和未加权分析进行细致检验可以提供有用的信息。在表 6.7 中，加权后和未加权的系数估计值类似。我们发现加权对截距项的影响要大于对系数的影响。表 6.7 所示的分析是用 logistic 回归分析数据的一个简单的演示，并没有特别考虑对合适模型的选择。我们并没有对表 6.8 中次序 logistic 回归模型和表 6.9 中多类别 logistic 回归模型进行类似的、没有使用权重和抽样设计的、基于模型的分析，因为我们并没有设定出一个合适的包含所有相关自变量的模型。

综上所述，复杂调查数据同时需要基于模型和基于设计的分析。基于设计的方法产生近乎无偏的估计值或者相关性，但标准误差可能是低效的。基于模型的方法需要在选择模型时作出假设，而错误的假设会导致相关性和标准误差的有偏估计。

# 第7章

# 总　结

在本书中,我们讨论了调查数据分析存在问题的各方面以及用于处理复杂抽样设计的使用所带来的问题的方法。重点关注的是对各种问题的理解和处理办法的逻辑,而不是技术操作指导。我们还介绍了准备复杂调查数据分析的实用指引,阐释了某些可用于执行各种分析的软件的使用。用于复杂调查分析的软件现在很容易找到,而且随着个人计算机计算能力的增强,许多复杂的分析方法也能很容易地执行。然而,数据分析者需要明确设计方法、对某些分析创建复合权重和对调查分析选择合适的检验统计值。因此,使用者应该对抽样设计以及相关的分析问题有很好的理解。

虽然对这些议题所介绍的材料主要是针对调查数据分析者,但我们希望这部分内容能同时激励调查设计者和数据制造者更加关注调查数据使用者的需要。调查数据最初的收集仅仅是为了记数,随着这些数据越来越多地用于分析性研究,调查设计者必须考虑把某些与设计相关的信息包括在数据之中,以实现更加合适的分析,同时减轻数据使用者的负担。除了提供阶层和抽样单位的代码,数据制造者还应该创建与设计相应的抽样权重,甚至还要包括结合了无应答和事后分层调整的复合权重。

最后，我们必须指出，我们一直是站在基于设计的统计推论立场之上的，同时也简要介绍了另一种方法，即基于模型的推论。这两种方法各有优势。简单来说，基于模型的推论假定样本是从概念上的超总体中得到的一组便利的观察值。设定好的模型下的总体参数是主要关注点，而抽样选择方案则没有推论那么重要。因此，在这里抽样设计的角色变得不重要了，统计估计使用的是这个设定好的模型下的预测方法。自然，如果模型设定是错误的，那么这种估计就会产生偏差，而且即使在大样本中偏差也可能是巨大的。基于设计的推论要求把抽样设计考虑在内，这是传统的方法。有限总体是主要关注点，分析目标在于在重复的抽样中找到设计无偏的估计值。

我们相信，当推论基于抽样数据时抽样设计是非常重要的，相比那些更具可预测性的物理现象，社会现象中的描述尤其重要。同时，我们需要对模型的恰当性进行评估，分析性模型的作用必须在所有数据分析中得到重视。同时使用设计和模型的推论相比只单独使用其中一种的推论可能更加成功。我们还相信，这两种不同的方法是互补的，联合使用两者会有益处。Brewer(1995、1999)和 Sundberg(1994)也陈述过类似的观点。[7]我们既不能忽略那些在社会调查设计中普遍存在的实际性问题，也不能忽略那些一直持续的无应答及其他引起非抽样误差的问题。这些实际性问题和社会科学研究中实质理论的现状迫使我们现在只能更多地依赖传统的方法。基于模型的方法应该在消除调查设计和分析间的鸿沟这个问题上提供额外的思考方法。

已有大量的理论和实际文献探讨了复杂调查数据的分

析问题。除了本书所列的参考文献,另有关于本书所讨论问题及相关议题的、更深入的、独立成册的分析(Korn & Graubard, 1999; Lehtonen & Pahkinen, 1995; Skinner、Holt & Smith, 1989)。在对复杂调查数据进行二手分析时,分析者应该意识到这种分析很可能会遭到抽样设计局限甚至抽样设计错误所导致的严重破坏。掌握有限抽样信息的分析者很有必要向有经验的专业调查技术员进行咨询。

# 注释

[1] 比率估计的方法用于估计两个变量的总体比率(比如,水果重量与产生果汁多少的比率)。它还用于获得某个变量的更加准确的估计值(比如现在的收入 $y$),通过构建其与另一个密切相关的变量的比率(比如上次普查时的收入 $x$)。这个样本比率($y/x$ 或收入的变化)接着被应用到上次普查时的收入来获得现在收入的估计值,它比没有使用辅助变量进行的估计更加准确。详细介绍见 Cochran(1977:第 6 章)。

[2] Holt 和 Smith(1979)把事后分层描述为一种稳健估计技术。依据条件分布,他们发现自加权后的样本均值通常是有偏差的,而事后分层则针对极端的样本构造为更准确的估计提供了保护。他们认为在现有的状况中,事后分层在抽样调查中没有得到充分应得的考虑。此外,事后分层还可以针对由抽样选择中的无应答和其他问题带来的任何异常现场为更准确的估计提供保护。

[3] SUPER CARP 和 PC CARP 可从爱荷华州立大学统计实验室获得,而且对偏统计的用户很有用。美国人口普查局的 Fay 博士使 CPLX 变得可用,它对用调整后的 BRR 以及折叠法进行离散多变量分析很有帮助(Fay, 1985)。CENVAR 和 VPLX 程序现在可从美国人口普查局找到。由 CDC(美国疾病控制和预防中心)开发的用于流行病学和统计学分析的 Epi Info 系统包括了用于复杂调查数据分析的 CSAMPLE 程序。密歇根大学社会研究所的 OSIRIS 统计软件系统包含了一些用于描述性统计和回归分析的程序(只对大型电脑开放)。还有其他用于特殊调查项目如世界生育率调查(CLUSTERS)的程序。另外有两个用于调查数据分析的 SAS 宏系列程序,包括 GES(可从加拿大统计局获得)和 CLAN(可从瑞典统计局获得)。

[4] 美国国民健康与营养调查(NHANES Ⅲ)是由美国国家卫生统计中心(NCHS)组织的一系列持续的调查,用以评估美国人口的健康与营养状况。NHANES 已经完成了好几轮调查。NHANES Ⅰ是在 1971—1973 年开展的,NHANES Ⅱ完成于 1976—1980 年,而 NHANES Ⅲ则在 1988—1994 年进行。1982—1984 年间他们还组织了一次专门针对西班牙裔人口的调查(西班牙 NHANES)。NHANES 已经变成一个持续的调查项目,数据现在每两年公开一次(1999—2000, 2001—2002, 2003—2004)。NHANES 从数目庞大的被访者中,通过个体访问和包括诊断检测与其他临床实践中用到的程序的健康检查,收集各种与健康相关的信息(S. S. Smith, 1996)。NHANES 的设计比较复杂,用以

适应经费和调查要求的实际限制,最终得到了一个包含家庭中符合条件个人的分层多阶段概率集群样本(NCHS, 1994)。其初级抽样单位(PSU)是县或相邻县的小组群,随后的分层抽样单位包括普查小区、家庭集群、家庭和符合条件的个人。对学前儿童、老人或者穷人采取了过取样,以提供足够的这些子群体的人数。公开的微观数据库所包含的抽样权重是扩展权重(根据无应答和事后分层而调整过的选择概率的倒数)。NHANES Ⅲ分两个阶段执行。多阶段抽样设计得到了 89 个抽样地区,这些地区接着又随机分成两组:其中 44 个点在 1988—1991 年间调查,剩下的 45 个在 1991—1994 年间调查。每个阶段的样本都可视为独立样本,而联合样本则可用于大型分析。

[5] 热卡填补(就近填补)的方法借用数据库中其他观察值的取值。有许多不同的选择捐赠观察值的方法。通常是按照选中的人口特征变量及其他变量如阶层和 PSU 等对数据进行排序来建立填补单元。然后从与具缺失值的观察值处于同一个单元的其他观察之中选取捐赠者。通过输入个体取值而非均值,这种方法在很大程度上避免对方差的低估。这种方法广泛用于美国人口普查局及其他调查机构。详情请见 Levy 和 Lemeshow(1999:409—411)。

[6] 自由度为 1 的 Wald 统计值基本上等于均值为 0 的正态变量的平方除以其标准差。对于涉及自由度大于 1 的假设,Wald 统计值是正态变量的平方的矩阵扩展。

[7] Brewer 和 Mellor(1973)精辟地阐述了基于设计和基于模型的方法之间的差别。Brewer(1995)充分地描述了不同方法的联合使用,这种联合使用方法进一步被应用到分层相对分层对称抽样的例子中(Brewer, 1999)。Sundberg(1994)结合方差估计表达了类似观点。Graubard 和 Korn(2002)回顾和阐释了从基于设计的抽样设计转到基于模型的抽样设计的优点以及抽样推论的优点。他们指出,在分层抽样中,把 FPC 因素从标准的基于设计的方差公式中去掉,以获得合适的方差公式作基于模型的推论是不充分的。在集群抽样中,标准的基于设计的方差公式会严重低估基于模型的变异情况,就算在最后(抽样)单位的抽样比例很小的情况下也如是。他们得出结论认为,基于设计的推论是一个有效的、合理不受模型干扰的方法,可对有限总体参数作出推论,但他们建议对基于设计的方差估计进行简单的修正,以对超总体参数做一些模型假设情况下的推论,而这也常常引起主要的科学关注。

# 参考文献


Aldrich. J. H. , & Nelson, F. D. (1984). *Linear probability, logit, and probit models* (Quantitative Applications in the Social Sciences, 07—045). Beverly Hilis, CA: Sage.

Alexander. C. H. (1987). A model-based justification for survey weights. *Proceedings of the Section of Survey Research Methods* (American Statistical Association), 183—188.

Bean, J. A. (1975). *Distribution and properties oj variance estimation for complex multistage probability samples* (Vital and Health Statistics, Series 2[65]). Washington, DC: National Center for Health Statistics.

Binder. D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, *51*, 279—292.

Brewer, K. R. W. (1995). Combining design-based and model-based infereoce. In B. G. Cox. D. A. Binder, B. N. Chinnappa. A. Christianson, M. J. Colledge, & P. S. KOll(Ed,.), *Business Survey methods*, 589—606. New York: John Wiley.

Brewer, K. R. W. (1999). Design-based or prediction-based inference? Stratified random vs. stratified balanced sampling. *International Statistical Review*, *67*, 35—47.

Brewer, K. R. W. , & Mellor, R. W. (1973). The effect of sample structure on analytical surveys. *Australian Journal of Statistics*, *15*, 145—152.

Brick, J. M. , & Kalton. G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, *5*, 215—238.

Bryk. A. S. , & Raudenbush. S. W. (J992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park. CA: Sage.

Chambless, L. E. , & Boyle, K. E. (1985). Maximum likelihood methods for complex sample data: Logistic regression and discrete proportional hazards models. *Communications in Statistics-Theory and Methods*, *14*, 1377—1392.

Chao, M. T. , & Lo, S. H. (1985). A bootstrap method for finite populalions. *Sankhya*, *47*(A), 399—405.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley.

Cohen, S. B. (1997). An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. *The American Statistician*, *51*, 285—292.

Davis, J. A., & Smith, T. W. (1985). *General Social Survey*, 1972—1985: *Cumulative codebook* (NORC edition). Chicago: National Opinion Research Center, University of Chicago and the Roper Center, University of Connecticut.

DeMaris, A. (1992). *Logit modeling* (Quantitative Applications in the Social Sciences, 07—086), Thousand Oaks, CA: Sage.

Deming. W. E. (1960). *Sample design in business research*. New York: John Wiley.

DuMouchel, W. H., & Duncan, G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, *78*, 535—543.

Durbin. J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika*, *46*, 477—480.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1—26.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Eliason. S. R. (1993). *Maximum likelihood estimation: Logic and practice* (Quantitative Applications in the Social Sciences, 07—096). Beverly Hills, CA: Sage.

Eltinge, J. L., Parsons, V. L., & Jang, D. S. (1997). Differences between complex-design-based and II D-based analyses of survey data: Examples from Phase I of NHANES III. *Stats*, *19*, 3—9.

Fay, R. E. (1985). A jackknife chi-square test for complex samples. *Journal of the American Statistical Association*, *80*, 148—157.

Flyer, P., & Mohadjer, L (1988), *The WesVar procedure*. Rockville, MD: Westat.

Forthofer. R. N., & Lehnen. R. G. (1981). *Public program analysis: A categoricai data approach*. Belmont, CA: Lifetime Learning Publications.

Frankel, M. R. (1971). *Inference from survey samples*. Ann Arbor: Insti-

tute of Social Research, University of Michigan.

Fuller, W. A.(1975). Regression analysis for sample surveys, *Sankhya*, *37* (C), 117—132.

Fuller, W. A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology*, *10*, 97—118.

Goldstein, H., & Silver, R. (1989). Multilevel and multivariate models in survey analysis. In C. J. Skinner, D. Holt, & T. M. F. Smith(Eds.), *Analysis of complex survey data* (pp. 221—235). New York: John Wiley.

Goodman, L. A. (1972). A general model for the analysis of surveys. *American Journal of Sociology*, *77*, 1035—1086.

Graubard, B. I., & Korn, E. L(1996), Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research*, *5*, 263—281.

Graubard, B. I., & Korn, E. L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, *17*, 73—96.

Grizzle, J. E., Starmer, C F., & Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics*, *25*, 489—504.

Gurney, M., & Jewett, R. S. (1975). Constructing orthogonal replications for variance estimation. *Journal of the American Statistical Association*, *70*, 819—821.

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Sratistical Association*, *78*, 776—807.

Heitjan, D. F. (1997). Annotation: What can be done about missing data? Approaches to imputation. *American Journal of Public Health*, *87* (4), 548—550.

Hinkins, S., Oh, H. L., & Scheuren, F. (1994). Inverse sampling design algorithms. *Proceedings of the Section on Survey Research Methods* (American Statistical Association), 626—631.

Holt, D., & Smith, T. M. F. (1979). Poststratification. *Journal of the Royal Statistical Society*. *142*(A), 33—46.

Holt, D., Smith, T. M. F., & Winter, P. D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society*, *143*(A), 474—487.

Horton, N. J., & Lipsitz, S. R. (2001). Multiple imputation in practice:

Comparison of software packages for regession models with missing variables. *The American Statistician*, *55*(3), 244—254.

Hosmer, D. W., & Lemeshow. S. (1989). *Applied logistic regression*. New York: John Wiley.

Judkins, D. R. (1990). Fay method of variance estimation. *Official Statistics*, *6*(3), 233—239.

Kalton, G. (1983). *Introduction to survey sampling* (Quantitative Applications in the Social Sciences. 07—035). Beverly Hills, CA: Sage.

Kalton, G., & Kasprszky, D, (1986), The treatment of missing survey data. *Survey Methodology*, *12*(1), 1—16.

Kendall, P. A.. & Lazarsfeld, P. F. (1950). Problems of survey analysis. In R. K. Merton & P. F. Lazarsfeld (Eds.), *Continuities in social research: Studies in the scope and method of "The American soldier."* New York: Free Press.

Kiecolt, K. J., & Nathan, L. E. (1985). *Secondary analysis of survey data* (Quantitarive Applications in the Social Sciences, 07—053). Beverly Hills, CA: Sage.

Kish. L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, *44*, 380—387.

Kish, L. (1965). *Survey sampling*. New York: John Wiley.

Kish, L., & Frankel, M. R. (1974). Inferences from complex samples. *Joumal of the Royal Statistical Society*, *36*(B), 1—37.

Knoke, D., & Burke, P. J. (1980). *Log-linear models* (Quantitative Applications in the Social Sciences, 07—020). Beverly Hills, CA: Sage.

Koch, G. G., Freeman, D. H.. & Freeman, J. L. (1975). Strategies in the multivariate analysis of data from complex surveys. *international Statistical Review*, *43*, 59—78.

Konijn. H. (1962). Regression analysis in sample surveys. *Journal of the American Statistical Association*, *57*, 590—605.

Korn, E. L., & Graubard, B. I. (1995a). Analysis of large health surveys: Accounting for the sample design. *Journal of the Royal Statistical Society*, *158*(A), 263—295.

Korn, E. L., & Graubard, B. I. (1995b). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*, *49*, 291—295.

Korn. E. L. , & Graubard. B. I. (1998). Scatterplots with survey data. *The American Statistician*, *52*, 58—69.

Korn. E. L. , & Graubard, B. I. (1999). *Analysis of health surveys*. New York: John Wiley.

Korn. E. L. , & Graubard, B. I. (2003). Estimating variance components by using survey data. *Journal of the Royal Statistical Society*, *8*(65, pt. 1), 175—190.

Kott, P. S. (1991). A model-based look at linear regression with survey data. *The American Statisrician*, *45*, 107—112.

Kovar. J. G. , Rao, J. N. K. , & Wu, C. F. J. (1988), Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, *16*(Suppl. ), 25—45.

Krewski, D. , & Rao, J. N. K. (1981), Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, *9*, 1010—1019.

LaVange, L. M. , Lafata. J. E,. Koch, G, G. , & Shah. B. V. (1996). Innovative strategies using SUDAAN for analysis of health surveys with complex samples. *Statistical Methods in Medical Research*, *5*, 311—329.

Lee, E. S. , Forthofer. R, N. , Holzer, C. E. , & Taube. C. A. (1986). Complex survey data analysis: Estimation of standard errors using pseudo-strata. *Journal of Economic and Social Measurement*, *14*, 135—144.

Lee, E. S. , Forthofer. R. N. , & Lorimor, R. J. (1986). Analysis of complex sample survey data: Problems and strategies. *Sociological Methods and Research*, *15*, 69—100.

Lee, K. H. (1972). The use of partially balanced designs for the half-sample replication method of variance estimation. *Journal of the American Statistical Association*, *67*, 324—334.

Lehtonen. R. , & Pahkinen. E. J. (1995). *Practical methods for design and analysis of complex surveys*. New York: John Wiley.

Lemeshow. S. , & Levy. P. S. (1979). Estimating the variance of ratio estimates in complex surveys with two primary sampling units per stratum. *Journal of Statistical Computing and Simulation*, *13*, 191—205.

Levy, P. S. , & Lemeshow, S. (1999). *Sampling of populations: Methods and applications*, New York: John Wiley.

Levy. P. S. , & Stolte. K. (2000). Statistical methods in public health and epidemiology: A look at the recent past and projections for the future. *Statistical Methods in Medical Research*, *9*, 41—55.

Liao. T. F. (1994). *Interpreting probability models: Logit, probit, and other generalized linear models* (Quantitative Applications in the Social Sciences. 07—101). Beverly Hills, CA, Sage.

Little. R. J. A. , & Rubin. D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley.

Little. R. J. A. , & Rubin. D. B. (2002). *Statistical analysis with missing data* (2nd ed. ). New York: John Wiley.

Lohr. S. L. (1999). *Sampling: Design and analysis*. New York: Duxbury.

McCarthy. P. J. (1966). *Replication: An approach to the analysis of data from complex surveys* (Vital and Health Statistics. Series 2[14]). Washington, DC: National Center for Health Statistics.

Murthy. M. N. , & Sethi, V. K. (1965). Self-weighting design at tabulation stage. *Sankhya*, *27*(B), 201—210.

Nathan, G. , & Holt, D. (1980). The effects of survey design on regression analysis. *Journal of the Royal Statistical Society*, *42*(B), 377—386.

National Center for Health Statistics(NCHS). (1994). *Plan and operation of the Third National Health and Nutrition Examination Survey*, 1988—1994(Vital and Health Statistics. Series 1[32]). Washington, DC: Government Printing Office.

Nielsen, S. F. (2003). Proper and improper multiple imputation. *International Statistical Review*, *71*, 593—627.

Nordberg, L. (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics*, *5*, 223—239.

Pfeffermann. D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, *61*, 317—337.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, *5*, 239—261.

Pfeffennann, D. , & Homes, D. J. (1985). Robustness considerations in the choice of method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society*, *148*(A), 268—278.

Pfeffennann. D. , & Nathan, G. (1981). Regression analysis of data from a cluster sample. *Journal of the American Statistical Association*, *76*, 681—689.

Plackett R. L. , & Burman, P. J. (1946). 'The design of optimum multi-factorial experiments. *Biometrika*, *33*, 305—325.

Quenouille, M. H. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society*, *11*(B), 68—84.

Rao, J. N. K. , Kovar, J. G. , & Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary infonnation. *Biometrika*, *77*, 365—375.

Rao, J. N. K. , & Scott, A. J. (1984). On chi-square tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, *12*, 46—60.

Rao, J. N. K. , & Wu, C. F. J. (1988). Resampling inference with complex: survey data. *Journal of the American Statistical Association*, *83*, 231—241.

Rao, J. N. K. , Wu. C. F. J. , & Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, *18*(*3*), 209—217.

Roberts. G. , Rao. J. N. K. , & Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, *74*, 1—12.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, *57*, 377—387.

Royall. R. M. (1973). *The prediction approach to finite population sampling theory: Application to the hospital discharge survey* (Vital and Health Statistics, Series 2[55]). Washington, DC: National Center for Health Statistics.

Rust, K. F. , & Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, *5*, 283—310.

Sarndal. C. E. (1978). Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics*, *5*, 25—52.

Sarndal. C. E. , Swensson, B. , & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Shah, B. V. , Holt, M. H. , & Folsom, R. E. (1977). Inference about regression models from sample survey data. *Bulletin of the International Statistical Institute*, *47*, 43—57.

Sitter, R. R. (1992). Resampling procedure for complex survey data. *journal of the American Statistical Association*, *87*, 755—765.

Skinner, C. J. , Holt, D. , & Smith. T. M. F. (Eds. ). (1989). *Analysis of complex survey data*. New York: John Wiley.

Smith, S. S. (1996). The third National Health and Nutrition Examination Survey: Measuring and monitoring the health of the nation. *Stats*, *16*, 9—11.

Smith, T. M. F. (1976). The foundations of survey sampling: A review. *Journal of the Royal Statistical Society*, *139*(A), 183—204.

Smith. T. M. F. (1983). On the validity of inferences on non-random samples. *Journal of the Royal Statistical Society*, *146*(A), 394—403.

Sribney, W. M. (1998). Two-way contingency tables for survey or clustered data. *Stata Technical Bulletin*, *45*, 33—49.

Stanek. E. J. , & Lemeshow, S. (1977). The behavior of balanced half-sample variance estimates for linear and combined ratio estimates when strata are paired to form pseudo strata. *American Statistical Association Proceedings*: *Social Statistics Section*, 837—842.

Stephan, F. F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association*, *43*, 12—39.

Sudman. S. (1976). *Applied sampling*. New York: Academic Press.

Sugden. R. A. , & Smith, T. M. F. (1984). Ignorable and informative designs in sampling inference. *Biometrika*, *71*, 495—506.

Sundberg. R. (1994). Precision estimation in sample survey inference: A criterion for choice between various estimators. *Biometrika*, *81*, 157—172.

Swafford, M. (1980). Three parametric techniques for contingency table analysis: Non-technical commentary. *American Sociological Review*, *45*, 604—690.

Tepping, B. J. (1968). Variance estimation in complex surveys. *American Statistical Association Proceedings*, *Social Statistics Section*, 11—18.

Tukey. J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, *29*, 614.

Tukey, J. W. (1986). Sunset salvo. *The American Statistician*, *40*, 72—76.

U. S. Bureau of the Census. (1986, April). *Estimates of the population of the United States*, *by age*, *sex*, *and race*. 1980 to 1985(Current Population Reports. Series P—25, No. 985). Washington, DC: Author.

Wolter, K. M. (1985). *Introduction to variance estimation*. New York:

Springer-Verlag.

Woodruff. R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statstical Association*, *66*, 411—414.

Zhang. P. (2003). Multiple imputation: Theory and method. *International Statistical Review*, *71*, 581—592.

# 译名对照表

| | |
|---|---|
| approximation | 近似值 |
| attrition-adjusted | 损耗调整 |
| attrition rate | 损耗率 |
| auxiliary variable | 辅助变量 |
| balanced repeated replication | 对称重复抽样 |
| body mass index (BMI) | 体重指数 |
| bootstrap method | 自主抽样法 |
| circular systematic sampling scheme | 循环系统抽样方案 |
| cluster sampling | 集群抽样 |
| collapsed strata | 折叠层 |
| confidence interval | 置信区间 |
| contingency table | 列联表 |
| covariance | 协方差 |
| Cox proportional hazards model | Cox 比例风险模型 |
| cross-sectional surveys | 截面调查 |
| cross-tabulation analysis | 跨列表分析 |
| design effect | 设计效应 |
| design-weighted least squares (DWLS) | 设计加权最小二乘法 |
| disproportionate stratified sample design | 不成比例分层抽样设计 |
| effect coding | 效果编码 |
| expansion estimation | 扩展估计 |
| expasion weight | 扩展权重 |
| extreme value | 极端值 |
| finite population correction factor | 有限总体校正因素 |
| first-order derivative | 一阶导数 |
| follow-up survey | 追踪调查 |
| generalized linear models | 广义线性模型 |
| general social survey(GSS) | 综合社会调查 |
| goodness of fit | 拟合优度 |
| host sample | 主样本 |
| hot deck imputation | 热卡填补(就近填补) |

| | |
|---|---|
| independent and identically distributed | 独立同分布 |
| indicator variable | 指示变量 |
| intraclass correlation coefficient(ICC) | 组内相关系数 |
| inverse sampling design algorithm | 逆抽样设计演算法 |
| item nonresponse | 选项无应答 |
| jackknife repeated replication | “折叠式”重复抽样 |
| joint inclusion probabilities | 联合被选中概率 |
| linear combination | 线性组合 |
| linearization method | 线性化方法 |
| log odds | 对数比 |
| logistic regression | logistic 回归 |
| log-likelihood | 对数似然 |
| log-linear regression | 对数线性回归 |
| longitudinal data | 纵向数据 |
| main effects model | 主效应模型 |
| mean imputation | 均值填补 |
| mean square error | 均方差 |
| measurement error | 测量误差 |
| median-trace plot | 中位数标示图 |
| missing value | 缺失值 |
| multicollinearity | 多重共线性 |
| multinomial logistic regression | 多类别 logistic 回归 |
| multiple imputation | 多重填补 |
| multiple-layered nesting | 多层嵌套 |
| multistage cluster sampling | 多阶段集群抽样 |
| national opinion research center(NORC) | 全国民意研究中心 |
| nonprobability sampling | 非概率抽样 |
| null model | 零模型 |
| odds ratio | 比数比 |
| optimum(Neyman) allocation | 最优(内曼)分配 |
| ordered logistic regression | 次序 logistic 回归 |
| ordinary least squares(OLS) estimation | 普通最小二乘估计 |

| | |
|---|---|
| orthogonal matrix | 正交矩阵 |
| oversampling | 过取样 |
| paired selection design | 配对选择设计 |
| parameter | 参数 |
| partially balanced replicate | 部分对称复合样本 |
| point estimate | 点估计 |
| poisson regression | 泊松回归 |
| population mean | 总体均值 |
| population total | 全及总体 |
| poststratification adjustment | 事后分层调整 |
| PPS(probability proportional to size) sampling | PPS 抽样(按规模大小成比例概率抽样) |
| poverty index | 贫困指数 |
| presicion | 精确度 |
| primary sampling unit(PSU) | 初级抽样单位 |
| probability distribution | 概率分布 |
| probability sampling | 概率抽样 |
| propagation of variance | 方差传递法 |
| proportional hazard model | 比例风险模型 |
| proportional odds assumption | 成比例发生比假设 |
| proportional odds model | 成比例发生比模型 |
| pseudo-replication | 拟复合抽样 |
| pseudo-strata | 拟阶层 |
| quota sampling | 配额抽样 |
| random variable | 随机变量 |
| ratio estimate | 比率估计 |
| regression imputation | 回归填补 |
| relative weight | 相对权重 |
| repeated systematic sampling | 重复系统抽样 |
| replicate weights | 复合权重 |
| replicated sampling | 复合抽样 |
| response rate | 应答率 |

| | |
|---|---|
| robust estimation | 稳健估计 |
| sample design | 抽样设计 |
| sample mean | 样本均值 |
| sample weights | 抽样权重 |
| sampling fraction | 抽样比例 |
| sampling frame | 抽样框 |
| saturated model | 饱和模型 |
| scatterplot | 散点图 |
| selection bias | 选择偏差 |
| self-weighting | 自加权 |
| side-by-side boxplot | 对比箱形图 |
| simple random sampling | 简单随机抽样 |
| simple random sampling with replacement (SRSWR) | 简单有放回随机抽样 |
| simple random sampling without replacement (SRSWOR) | 简单无放回随机抽样 |
| simple two-stage cluster sampling | 简单二阶段集群抽样 |
| simultaneous equation | 联立方程 |
| standard error | 标准误差 |
| statistical inference | 统计推断 |
| stratified multistage cluster sampling | 分层多阶段集群抽样 |
| stratified random sampling | 分层随机抽样 |
| stratified sampling | 分层抽样 |
| survival analysis | 生存分析 |
| synthetic estimation | 综合估计 |
| systematic sampling | 系统抽样 |
| systolic blood pressure(SBP) | 收缩压 |
| target population | 目标总体 |
| Taylor series method | 泰勒级数法 |
| the best unbiased predictor | 最优无偏估计 |
| ultimate cluster approximation | 最终集群逼近法 |
| unequal probability design | 不等概率设计 |

| | |
|---|---|
| unequal-sized cluster | 不同规模集群 |
| unit nonresponse | 单元无应答 |
| variance-covariance matrix | 方差一协方差矩阵 |
| variance estimation | 方差估计 |
| weighted average | 加权平均数 |
| weighted fitting | 加权拟合 |
| weighted least-square | 加权最小二乘法 |
| weighted least-square modeling | 加权最小二乘建模 |
| weighted odds ratios | 加权后对数比 |
| weighted proportions | 加权后比例 |
| weighted sum | 加权总和 |